BB

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 715 298 B1

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention of the grant of the patent:
**06.09.2000 Bulletin 2000/36**

(51) Int Cl.[7]: **G10L 15/04**, G10L 15/28, G10L 15/14

(21) Application number: **95109575.1**

(22) Date of filing: **21.06.1995**

(54) **Reduction of search space in speech recognition using phone boundaries and phone ranking**

Verminderung des Suchraumes bei Spracherkennung unter Verwendung von Phonemgrenzen und Phonemklassen

Réduction de l'espace de recherche dans la reconnaissance de la parole sous utilisation des limites et classement des sons

(84) Designated Contracting States:
**DE FR GB**

(30) Priority: **30.11.1994 US 347013**

(43) Date of publication of application:
**05.06.1996 Bulletin 1996/23**

(73) Proprietor: **International Business Machines Corporation**
**Armonk, N.Y. 10504 (US)**

(72) Inventors:
- **Nahamoo, David**
  **White Plains, New York 10605 (US)**
- **Padmanabhan, Mukund**
  **Ossining, New York 10562 (US)**

(74) Representative: **Schäfer, Wolfgang, Dipl.-Ing.**
**IBM Deutschland**
**Informationssysteme GmbH**
**Patentwesen und Urheberrecht**
**70548 Stuttgart (DE)**

(56) References cited:
**EP-A- 0 313 975**
**EP-A- 0 387 602**
**EP-A- 0 424 665**



Printed by Jouve, 75001 PARIS (FR)

EP 0 715 298 B1

## Description

### TECHNICAL FIELD

**[0001]** The invention pertains to speech recognition, and in particular phone boundary detection in the acoustic input.

### TERMS

**[0002]** Symbol: Characterizing acoustic speech based on n features, acoustic speech is viewed in an n-dimensional acoustic space. The space is partitioned into regions, each of which is identified by an n-dimensional prototype vector. Each prototype vector is represented by a "symbol", such as a number or other identifier. Uttered speech may be viewed as successive "symbols".

**[0003]** Feneme (also Label): A symbol corresponding to a prototype vector, the symbol being defined based on features of sound occurring during a fixed interval of time. Sound may be characterized as having, for example, twenty features-the magnitude of each feature during a centisecond interval corresponding to a prototype vector component. Each prototype vector thus has a corresponding set of feature values for a centisecond interval. Based on the feature values generated during a centisecond interval, one prototype vector from a fixed set of prototype vectors is selected as the closest. With each prototype vector having a corresponding feneme (or label), the set of prototype vectors corresponds to an alphabet of fenemes (or labels). Sample fenemes are listed in Table 1-the first feneme 001 being defined as AA11. An acoustic processor examines uttered speech one interval after another and, based on which prototype vector is closest by some measure to the feature values, the feneme for the closest prototype vector is assigned to the interval. The feneme is distinguished from the well-known phoneme in that the former is based on feature values examined over a fixed interval of time (e.g., a centisecond) whereas the latter is based on a predefined set of basic phonetic sound units without regard to time limitations.

**[0004]** Markov Model (also probabilistic finite state machine): A sound event can be represented as a collection of states connected to one another by transitions which produce symbols from a finite alphabet. Each transition from a state to a state has associated with it a probability which is the probability that a transition t will be chosen next when a state s is reached. Also, for each possible label output at a transition, there is a corresponding probability. The model starts in one or more initial states and ends in one or more final states.

**[0005]** Phone: A unit of sound for which a Markov model is assigned. A first type of phone is phonetically based, each phoneme corresponding to a respective phone. A standard set of phonemes are defined in the International Phonetic Alphabet. A second type of phone is feneme-based, each feneme corresponding to a respective phone.

**[0006]** Polling: From a training text, it is determined how often each label occurs in each vocabulary word. From such data, tables are generated in which each label has a vote for each vocabulary word and, optionally, each label has a penalty for each word. When an acoustic processor generates a string of labels, the votes (and penalties) for each vocabulary word are computed to provide a match value. The process of tallying the votes is "polling".

**[0007]** In some known approaches to speech recognition, words are represented by phone-based Markov models and input speech which, after conversion to a coded sequence of acoustic elements or labels, is decoded by matching the label sequences to these models, using probabilistic algorithms such as Viterbi decoding.

### BACKGROUND

A. Overview of Speech Recognition

**[0008]** (1) Labeling of Speech Input Signal A preliminary function of this speech recognition system is the conversion of the speech input signal into a coded representation. This is done in a procedure that was described for example in "Continuous Speech Recognition with Automatically Selected Acoustic Prototypes Obtained by either Bootstrapping or Clustering" by A. Nadas et al, Proceedings ICASSP 1981, pp. 1153-1155.

**[0009]** In accordance with the Nadas et al conversion procedure, speech input is divided into centisecond intervals. For each centisecond interval, a spectral analysis of the speech input is made. A determination is then made as to which of a plurality of predefined spectral patterns the centisecond of speech input most closely corresponds. A "feneme" that indicates which spectral pattern most closely conforms to the speech input is then assigned to the particular centisecond interval. Each feneme, in turn, is represented as a distinct label.

**[0010]** A string of labels (or fenemes) thereby represents successive centiseconds of speech which, in turn, form words.

**[0011]** A typical finite set of labels is shown in Table 1 which is appended to this specification. It comprises about 200 labels each of which represents an acoustic element. It should be noted that these acoustic elements are shorter than the usual "phonemes" which roughly represent vowels or consonants of the alphabet, i.e., each phoneme would

correspond to a sequence of labeled acoustic elements.

**[0012]** An important feature of this labeling technique is that it can be done automatically on the basis of the acoustic signal and thus needs no phonetic interpretation. The unit which does the conversion from the acoustic input signal to a coded representation in the form of a label string is called an "acoustic processor".

**[0013]** (2) Statistical Model Representation of Words

The basic functions of a speech recognition system in which the present invention can be used will be described here briefly though several publications are also available which give more details of such a system, in particular F. Jelinek, "Continuous Speech Recognition by Statistical Methods", Proceedings IEEE, Vol. 64, 1976, pp. 532-576.

**[0014]** In the system, each word of the recognition vocabulary is represented by a baseform wherein the word is divided for recognition purposes into a structure of phones, i.e. phonetic elements as shown in FIG. 1. These phones correspond generally to the sounds of vowels and consonants as are commonly used in phonetic alphabets. In actual speech, a portion of a word may have different pronunciations as is indicated by the parallel branches in FIG. 1. The parallel branches which extend between nodes through which all such branches pass may alternatively be considered together as a "clink" or as separate conventional phones. The clink, as the principles of this invention apply, may be viewed as a substitute phonetic element for the phones discussed hereinbelow. The phones in turn, are represented by Markov models. Referring now to FIG. 2 sample Markov model for a phone is illustrated. For each phone there is a corresponding Markov model characterized by (a) a plurality of states (S0 . . . S4), (b) transitions (T1 . . . T10 ) between the states, and (c) label probabilities, each representing the likelihood that the phone will produce a particular label at a given transition. In one embodiment each transition in the Markov model has two hundred stored label probabilities associated therewith, each probability representing the likelihood that each respective label (of a set of 200 labels) is produced by the phone at a given transition. Different phones are distinguished in their respective Markov models by differences in the label probabilities associated with the various transitions. The number of states and transitions therebetween may differ but, preferably, these factors remain the same and the stored label probabilities vary.

**[0015]** In the Markov model of FIG. 2, a string of labels SX1-SX3-SX5-SH2 (taken from Table 2) has entered the phone model in the order shown. The probability of each label occurring at the transition at which it is shown (e.g. SX1 at transition T1) is determined based on the corresponding stored label probability. Phone models having the highest label probabilities for the labels in the string are the most likely phones to have produced the string.

**[0016]** While the labels in FIG. 2 suggest continuity from label to label along transition to transition-which enables a simple one-to-one alignment between string label and transition-the Markov model of FIG. 2 also permits other alignment as well. That is, the Markov model of FIG. 2 can determine that a phone is likely even where more labels, less labels, or even different labels are applied to the phone model. In this regard, besides transitions from one state to another, there are also transitions (T5, T6, T7) that go back to the same state that was just left. Furthermore, there are transitions (T8, T9, T10) that skip a neighbor state. The Markov model thereby provides that different pronunciations of a phone can be accommodated in the same basic Markov model. If, for example, a sound is stretched (slow speaker) so that the same acoustic element appears several times instead of only once as usual, the Markov model representation allows several transitions back to the same state thus accommodating the several appearances of the acoustic element. If, however, an acoustic element that usually belongs to a phone does not appear at all in a particular pronunciation, the respective transition of the model can be skipped.

**[0017]** Any possible path (Markov chain) from the initial state to the final state of the Markov model (including multiple occurrences of the turnback transitions, T5, T6 or T7) represents one utterance of the word (or phone), one acoustic element or label being associated with each transition.

**[0018]** In the present invention, label strings are "aligned" to Markov models by associating labels in the string with transitions in a path through the model; determining probabilities of each label being at the associated transition, on the basis of stored label probabilities set by previous experiences or training (as explained below). A chain of Markov models having the highest probability identifies the word that is to be selected as output.

**[0019]** The baseforms of the words and the basic Markov models of phones can be derived and defined in different ways, as described in the cited literature. Model generation may be done by a linguist, or the models can be derived automatically using statistical methods. As the preparation of the models is not part of the invention, it will not be described in more detail.

**[0020]** It should be mentioned that instead of representing words first by a sequence of Markov phone models, they could also be directly represented by Markov word models-as by a sequence of states and transitions that represent the basic string of acoustic elements for the whole word.

**[0021]** After structuring of the basic models that represent the words in a vocabulary, the models must be trained in order to furnish them with the statistics (e.g. label probabilities) for actual pronunciations or utterances of all the words in the vocabulary. For this purpose, each word is spoken several times, and the label string that is obtained for each utterance is "aligned" to the respective word model, i.e. it is determined how the respective label string can be obtained by stepping through the model, and count values are accumulated for the respective transitions. A statistical Markov model is formulated for each phone and thus for each word as a combination of phones. From the Markov model it

can be determined with what probability each of various different label strings were caused by utterance of a given word of the vocabulary. A storage table representing such a statistical Markov model is shown in FIG. 3 and will be explained in more detail in a later section.

**[0022]** For actual speech recognition, the speech signal is converted by the acoustic processor to a label string which is then "matched" against the existing word models. A specific procedure, the Viterbi Algorithm (described briefly in the above mentioned Jelinek paper and in more detail in a paper by G. D. Forney, "The Viterbi Algorithm", Proceedings, IEEE, Vol. 61, 1973, pp. 268-278) is used for this, and the result is a probability vector for each of a number of "close" words which may have caused the given label sequence. Then the actual output, i.e. the identification of a word that is selected as the recognition output, is determined by selecting the word whose probability is found to have the highest generated probability vectors.

**[0023]** The estimation of phone probabilities is an essential part of "the match". Typically, the recognition is carried out in a maximum likelihood framework, where all words in the vocabulary are represented as a sequence of phones, and the probability of a given acoustic feature vector, conditioned on the phone is computed (i.e. P (acoustic/phone)). The recognition process hypothesizes that a given word in the vocabulary is the correct word and computes a probabilistic score for this word as described above; this is done for all words in the vocabulary, subsequently, the acoustic score is combined with a score provided by a language model, and the word with the highest combined score is chosen to be the correct one.

**[0024]** The probability P (acoustic/phone) is equal to the probability that the current state of the Markov model for the phone produces the observed acoustic vector at the current time, and this probability is accumulated over several time frames till the cumulative product falls below a defined threshold, at which point it is hypothesized that the phone has ended and the next phone has started. In this technique, it is possible that in computing this score, frames that do not actually belong to the current phone are also taken into account while computing the score for the phone. This problem can be avoided if the beginning and end times of a phone are known with a greater level of certainty. A technique to estimate the boundary points is given in ["Transform Representation of the Spectra of Acoustic Speech Segments with Applications - I:General Approach and Speech Recognition", IEEE Transactions on Speech and Audio Processing, PP. 180-195, vol. 1, no. 2, April 1993], which is based on using the relative variation between successive frames, however it is quite expensive computationally, and is also constrained in terms of the extent of the acoustic context that it considers.

**[0025]** In some speech recognition systems, "the match" is carried out in two stages. The first stage of the decoder provides a short list of candidate words, out of the 20K vocabulary. Subsequently, detailed models of the words in this short list are used to match the word to the acoustic signal, and the word with the highest score is chosen. The process for determining the short list, called the fast match (See U.S. Patent 5263117 titled "Method and Apparatus for Finding the Best Splits in a Decision Tree for a Language Model"), organizes the phonetic baseforms of the words in the vocabulary in the form of a tree, and traverses down this tree, computing a score for each node, and discarding paths that have scores below a certain threshold. A path comprises of a sequence of phones, and often, the score for several phones has to be computed before a decision can be made to discard the path. In an earlier invention ("Channel-Bank-Based Thresholding to Improve Search Time in the Fast Match", IBM TDB pp. 113-114, vol. 37, No. 02A, Feb. 1994), a method was described, whereby, by observing the output of a channel-bank, a poor path could be discarded at a very early stage, thus saving the cost of computing the scores for the remaining phones on the path. In "Channel-Bank-Based Thresholding to Improve Search Time in the Fast Match", IBM TDB pp. 113-114, vol. 37, No. 02A, Feb. 1994 the channel-bank outputs were computed in a "blind" fashion, as no information was available about the start and end times of a phone in the acoustic label sequence. In this invention, we describe a method of computing the channel-bank outputs in a more intelligent fashion, that results in a reduction in the overall error rate and also reduces the computation time of the fast match.

**[0026]** Correspondingly, there is proposed a method as set out in claims 1 and 4, and an apparatus as set out in claims 6 and 9.

**[0027]** This invention proposes an alternative technique to predict phone boundaries that enables the use of an extended acoustic context to predict whether the current time is a phone boundary. The invention uses a non-linear decision-tree-based approach to solve the problem. The quantized feature vectors at, and in the vicinity of, the current time are used to predict the probability of the current time being a phone boundary, with the mechanism of prediction being a decision tree. The decision tree is constructed from training data by designing binary questions about the predictors such that the uncertainty in the predicted class is minimized by asking the question. The size of the class alphabet here is 2, and the technique of [L. Breiman, J.H. Friedman, R.A.Olshen, C.J. Stone, "Classification and Regression Trees", Wadsworth, Inc., 1984] is used to design the questions for each predictor.

**[0028]** The invention also describes a technique that can be used to further cut down the search space of the speech recognition system. Assuming that the phone boundaries are known, it is possible to compute the score for all phones in the segment between two phone boundaries, and compute the rank of the correct phone in this segment. Ideally, of course, the correct phone should be ranked first, and it should be possible to eliminate all phones other than the topmost

phone from the search space. However, in reality, due to ambiguities in the acoustic modelling, the vector-quantized acoustic feature vectors in the segment may not be representative of the sound or phone which was actually uttered in the segment.

**[0029]** Consequently, the rank of the correct phone could be quite poor in certain segments.

**[0030]** The invention also describes a decision-tree-based technique to predict the worst case rank of the correct phone between two hypothesized phone boundaries. Once this worst case rank is known, all the phones that are ranked below the worst case rank are eliminated from the search space of the recognizer, resulting in a large saving in computation. Note that the technique is independent of the method used to compute the score for a phone; typical schemes are (a) the usual Markov-model based computation (b) a channel-bank-based computation as described in ["Channel-Bank-Based Thresholding to Improve Search Time in the Fast Match", IBM TDB pp. 113-114, vol. 37, No. 02A, Feb. 1994] and (c) a decision-tree-based scoring mechanism, as described in [co-pending US Patent Application, D. Nahamoo, M. Padmanabhan, M.A. Picheny, P.S. Gopalkrishnan, "A Decision Tree Based pruning strategy for the Acoustic Fast Match, IBM Attorney Docket YO 996-059], or any alternative scoring mechanism.

**[0031]** The predictors used in the decision tree are, as before, the quantized acoustic feature vectors at, and in the vicinity of, the current time, and the predicted quantity is the worst case rank of the correct phone at the current time. The decision tree is constructed from training data by designing binary questions about the predictors, which are asked while traversing down the nodes of the decision tree. The questions are designed to minimize the uncertainty in the predicted class. Unlike the previous case of boundary estimation, however, the size of the class alphabet is equal to the number of phones, which is typically much larger than 2, and the technique outlined in ["Method and Apparatus for Ginding the Best Splits in a Decision Tree for a Language Model for a Speech Recognizer, U.S. Patent 5263117] is used to design the questions for each node.

**[0032]** The objective of the invention is to take the given vector-quantized feature vectors at the current time t, and the adjacent N time frames on either side, and devise two decision-trees. The first decision-tree should give the probability of the current frame being a phone boundary, and the second decision tree should give a distribution over all possible ranks that the correct phone can take at that time, from which the worst case rank of the current phone can be obtained.

**[0033]** A decision tree having true or false (i.e., binary) questions at each node and a probability distribution at each leaf is constructed. Commencing at the root of the tree, by answering a question at each node encountered and then following a first or second branch from the node depending upon whether the answer is "true" or "false", progress is made toward a leaf. The question at each node is phrased in terms of the available data (e.g., the words already spoken) and is designed to ensure that the probability distribution at the leaves provide as much information as possible about the quantity being predicted.

**[0034]** A principal object of the invention is, therefore, the provision of a method of designing and constructing a binary decision tree having true or false questions at each node starting from the root of the tree towards a leaf.

**[0035]** Another object of the invention is the provision of a method of constructing a binary-decision tree using questions phrased in terms of the available known data and designed to ensure that the probability distribution at the leaves maximize the information about the quantity being predicted.

**[0036]** A further object of the invention is the provision of a method of constructing a binary decision tree primarily for use in speech pattern recognition.

**[0037]** Further and still other objects of the invention will become more clearly apparent when the following description is read in conjunction with the accompanying drawings.

**[0038]** The invention incorporates the following features:

a) The boundary points of phones in the acoustic label sequence are estimated by using a decision tree and the adjoining labels, i.e., in the context of the labels adjacent on both sides of the current label, a decision is made as to whether the current label represents the boundary point between two phones. In the remainder of this disclosure, the term "segment" will be used to denote the time interval between two boundary points.

b) Based only on the labels in a segment, a score is computed for all possible phones, based on the probabilities obtained from the decision tree described in ("Channel-Bank-Based Thresholding to Improve Search Time in the Fast Match", IBM TDB pp. 113-114, vol. 37, No. 02A, Feb. 1994). As mentioned earlier, alternative scoring mechanisms could be used to compute the score for a phone. The phones are next ranked in accordance to their scores.

c) A decision is made that all phones above a certain rank are "good" phones that are possible in the time segment of interest, and that the phones below this threshold rank are "bad" phones that are not possible in the segment of interest. The threshold rank is not fixed but is a function of the label sequence in the current segment and the adjacent segment, and is obtained by using a decision tree. The decision is made on the basis of the label at the start of the segment and the adjacent labels on either side of this label.

d) To avoid errors due to the pruning, the number of candidate phones is now increased by using phone classes, i.e., from training data a list is made for each phone, of the phones that are confusable with it. When decoding, for

every "good" phone obtained in step (c), all phones in the confusion class of the "good" phone are also designated as "good" phones.

e) An alternative to eliminating all "bad" phones from the search is to penalize the score for these bad phones in all subsequent computations in the fast match. All this is precomputed before the actual fast match.

**[0039]** The implementation of the algorithm in the decoder takes the following steps:
Given a sequence of labels, the following precomputation is done before the fast match: the phone probabilities are first computed from a decision tree as described in ("Channel-Bank-Based Thresholding to Improve Search Time in the Fast Match", IBM TDB pp. 113-114, vol. 37, No. 02A, Feb. 1994). Subsequently, the boundary points of phones in the acoustic label sequence are determined by using the first decision tree described above, and the ranks of different phones are computed within all segments, based on the probabilities obtained from the decision tree of ("Channel-Bank-Based Thresholding to Improve Search Time in the Fast Match", IBM TDB pp. 113-114, vol. 37, No. 02A, Feb. 1994). Then the threshold rank that should be applied in every segment is obtained by traversing down the second decision tree described above. The phones ranked above the threshold, and the phones in union of their confusion classes, are then designated as "good" phones, and the remainder as "bad" phones. The probabilities for the "bad" phones in the given segment are then penalized. This penalization is done both on the phone probabilities obtained from the decision tree of ("Channel-Bank-Based Thresholding to Improve Search Time in the Fast Match", IBM TDB pp. 113-114, vol. 37, No. 02A, Feb. 1994), and on the acoustic fast match probabilities.

**[0040]** Subsequently, the fast match tree is pruned using the modified probabilities above, using the techniques described in ("Channel-Bank-Based Thresholding to Improve Search Time in the Fast Match", IBM TDB pp. 113-114, vol. 37, No. 02A, Feb. 1994,"Transform Representation of the Spectra of Acoustic Speech Segments with Applications -I:General Approach and Speech Recognition", IEEE Transactions on Speech and Audio Processing, PP. 180-195, vol. 1, no. 2, April 1993).

**[0041]** Hence, the training data used for the construction of the decision tree consists of sets of records of 2N+1 predictors (denoted by the indices -N,...0,...N) and the class associated with index 0, (which is assumed to be known). The associated class, in the case of the first decision tree is a binary record that specifies whether or not the frame at index 0 is a phone boundary. The associated class, in the case of the second decision tree is the rank of the correct phone at index 0. The alphabet size of each predictor is in the hundred's, and the class alphabet size is either 2 in the case of the first decision tree, or typically 50 or so in the case of the second decision tree. The invention uses the technique described below to construct the two decision trees (note that the two trees are constructed independently of one another).

**[0042]** The invention uses a successive data partitioning and search strategy to determine the questions of the decision tree. Starting with all the training data at the root of the tree, the invention chooses one of the 2N+1 predictors and partitions the alphabet of the predictor into two non-overlapping sets. Subsequently, for all the training records at the current node, if the value of the chosen predictor lies in the first set, the record is assigned to the first set, otherwise it is assigned to the second set. Hence, the training data at the current node is distributed between two child nodes on the basis of the set membership of the selected predictor. The predictor and the partitioning of the alphabet are chosen in such a way that after the training data is partitioned as described above, the uncertainty in the predicted class is minimized. The procedure is repeated for each child of the current node, till the class uncertainty at a node (quantified by the entropy of the class distribution at the node) falls below a certain level, or till the amount of training data at a node falls below a certain level. After the tree is constructed, the class distribution at the terminal nodes of the tree is available, and is stored along with the questions of the tree.

**[0043]** For the case of the first decision tree, the stored quantity is simply the probability that the node is a phone boundary. For the case of the second decision tree, the quantity available at the nodes of the tree is a distribution over all possible ranks that the correct phone can take. This distribution is converted to a single number, a worst case rank, such that the probability that the rank of the correct phone is better than the worst case rank is stored at the node of the decision tree.

**[0044]** For the case of a single predictor and a class, Nadas and Nahamoo [U.S. Patent 5236117] describe a technique to find the best binary question that minimizes the uncertainty in the predicted class. At the current node, this technique is applied independently to each of the 2N+1 predictors, and the best question for this predictor is determined. Subsequently, the best one among the 2N+1 predictors is determined as the one that provides the maximum reduction in class uncertainty and the question at the current mode is formulated as the best question for this prediction. Alternatively, the question at a node could also be made more complex, such that it depends on more than one predictor, or an inventory of fixed complex questions could be used, and the best question chosen as the one in this inventory that provides the maximum reduction in class uncertainty.

**[0045]** It is another object of the invention to describe means whereby the above described decision tree can be used in a speech recognizer. During recognition, the first decision tree is traversed till it reaches one of the terminal nodes, and the probability of the current time being a phone boundary is obtained from the terminal node of the decision

tree. This is compared to a predetermined threshold, and if it is larger than the threshold, the current time is hypothesized to be a boundary point. Subsequently, the second decision tree is traversed for all time frames between two hypothesized phone boundaries, and the worst case rank of the correct phone is obtained from the terminal node of the decision tree, for all these time frames. The worst of these worst case ranks is taken to be the worst case rank of the correct phone in that segment. Subsequently, the score for all phones is computed on the basis of that segment, and the phones are ranked according to their scores. Then the phones that are ranked below the worst case rank are discarded from the search, thus making up a shortlist of allowed phones for every segment between two hypothesized phone boundaries. This list may also be augmented further by considering phones that are confusable with each other, and by including every element of a "confusable" list in the short list whenever any one element in the confusable list is ranked above the worst case rank.

**[0046]** This information is used in the maximum likelihood framework to determine whether to carry out a match for a given word, by constraining the search space of the recognizer to the shortlist, rather than the space of the entire alphabet. Before carrying out the match for a given phone in a word, the above defined shortlist is checked to see if the phone can possibly occur at the given time, and if the phone does not occur in the shortlist, then the match for the current word is discarded.

**[0047]** The method and apparatus according to the invention are advantageous because (a) they provide a fast and accurate way of estimating phone boundaries, by enabling the match for a phone to be done within well defined boundaries thus leading to better accuracy (b) they provide a fast and accurate means of estimating the rank boundaries of the correct phone without requiring any knowledge about the identity of the correct phone, and thus enable the creation of a shortlist of allowed phones, which helps in greatly cutting down the search space of the speech recognizer. Further, the overhead associated with traversing the two decision tree's is negligible, as the questions asked in the decision tree simply involve the set membership of the selected predictor.

Fig. 1 is an illustration of phonetic baseforms for two words;
Fig. 2 is a schematic representation of a Markov model for a phone;
Fig. 3 shows a partial sample of a table representing a statistical Markov model trained by numerous utterances.
Fig. 4 is a flow chart describing a procedure for constructing a decision tree to predict the probability distribution of a class at a given time, in accordance with the invention.
Fig. 5 is a schematic for constructing a decision tree.
Fig. 6. is a flow chart of an automatic speech recognition system using two decision trees.
Fig. 7 is a flow chart of an automatic speech recognition system using two decision trees.

**[0048]** Figure 4 is a flow chart depicting the procedure to construct a decision tree to predict a probability distribution on the class values at time t, given the quantized feature vectors at times t-N,t-N+1,...,t,...,t+N. For the purpose of explaining the working of the invention, the quantized feature vectors will henceforth be referred to as labels. The predictors used in the decision tree are the labels at times t-N,...,t...,t+N, represented as $l^{-N},...l^{0},...,l^{+N}$, and the predicted quantity is either a distribution over two classes as in the case of the boundary-detection decision tree, i.e., the probability that the time t is a phone boundary, or a distribution over all possible ranks of the correct phone at time t, as in the case of the rank-determining decision tree. The size of the class alphabet in the second case is equal to the size of the phone alphabet, denoted as P. The size of the label alphabet is denoted as L. Typically, P ranges from 50-100, and L is in the 100's; however, for the purpose of explaining the invention, we will assume that L=4, P=3, and N=1. We will represent these 4 predictor values as $l_1,l_2,l_3$, and $l_4$, and the 3 class values as $p_1,p_2$, and $p_3$. The technique described below uses the procedure of [1] to determine the binary partitioning of the predictor alphabet at a node of the decision tree, which is appropriate for the case of the rank-determining decision tree, where the number of classes is larger than 2. However, for the boundary-detection decision tree, where the number of classes is equal to 2, [U.S. Patent 5263117 titled "Method and Apparatus for Finding the Best Splits in a Decision Tree for a Language Model"] reduces to the simpler optimal strategy of [L. Breiman, J.H. Friedman, R.A.Olshen, C.J. Stone, "Classification and Regression Trees", Wadsworth, Inc., 1984].

**[0049]** The training data consists of a number of transcribed sentences, with the acoustic corresponding to each sentence being quantized into a sequence of labels. Further as the data is transcribed, it is also possible to assign a class value to every time frame.

**[0050]** If the event $l_p^k$ is defined as one where the value of the predictor $l^k$ is equal to $l_i$, and the class value is equal to p, then a confusion matrix is next created (Block 2), which enumerates the counts of all possible events $(l_i^k,p)$. The matrix has L rows, and P columns, and the entry corresponding to the $i^{th}$ row and the $j^{th}$ column represents the number of times the value of the predictor $l^k$ equalled $l_i$, when the class value equalled $p_j$, in the training data at the current node of the decision tree (at the root node, all the training data is used). These counts are then converted into joint probabilities by computing the sum of all entries in the matrix, and then dividing each entry of the matrix by this sum. As there are 2N+1 predictors, 2N+1 joint distribution matrixes can be created, one for each predictor. An example of

these joint distribution matrices is shown in Table 2 below, for the case of 3 predictors $I^{-1}$, $I^0$, and $I^{+1}$.

TABLE 2

| $I^{-1}$ | $p_1$ | $p2$ | $p_3$ |
|---|---|---|---|
| $l_1$ | 0.1 | 0.067 | 0.033 |
| $l_2$ | 0.067 | 0.167 | 0.033 |
| $l_3$ | 0.133 | 0.033 | 0.1 |
| $l_4$ | 0.033 | 0.067 | 0.167 |
| $I^0$ | $p_1$ | $p2$ | $p_3$ |
| $l_1$ | 0.133 | 0.05 | 0.033 |
| $l_2$ | 0.067 | 0.2 | 0.034 |
| $l_3$ | 0.1 | 0.034 | 0.067 |
| $l_4$ | 0.033 | 0.05 | 0.2 |
| $I^{+1}$ | $p_1$ | $p2$ | $p_3$ |
| $l_1$ | 0.117 | 0.05 | 0.033 |
| $l_2$ | 0.067 | 0.167 | 0.033 |
| $l_3$ | 0.116 | 0.05 | 0.1 |
| $l_4$ | 0.033 | 0.067 | 0.167 |

**[0051]** The class distribution at the current node and its entropy is computed and stored at this point. The class distribution is obtained by summing up the rows of any one of the 2N+1 joint distribution matrices, i.e.

$$Pr(p=p_k) = \sum_{j=1}^{4} Pr(l^k=l_j, p=p_k),$$

and the entropy of the class distribution is obtained as

$$H(p) = \sum_{i=1}^{3} -Pr(p=p_i) \log[Pr(p=p_i)].$$

**[0052]** For the considered example, the class distribution and its entropy is given in Table 3.The log in H(p) is base 2.

TABLE 3

| | $p_1$ | $p_2$ | $p_3$ |
|---|---|---|---|
| Pr | 0.333 | 0.334 | 0.333 |
| H(p) = 1.58 | | | |

**[0053]** In Block 3, we start with the joint distribution of the $k^{th}$ predictor, $I^k$, and the class p, and design a binary partitioning $SL^k_{opt}$, $\overline{SL^k_{opt}}$ of the values of the predictor $I^k$ using the method of [U.S. Patent 5236117 referenced above]. In other words, for each predictor, the predictor alphabet $[l_1,l_2,l_3,l_4]$ is partitioned into two complementary sets, $SL^k_{opt}$ and $\overline{SL^k_{opt}}$ (for example, $SL^k_{opt} = [l_1,l_2]$, *and* $\overline{SL^k_{opt}} = [l_3,l_4]$), with the criterion for the selection of the partition being the minimization of the class uncertainty. The entropy of the class distribution is used as a measure of the uncertainty. The details of this method are given in [U.S. Patent 5236117]. This process is carried out of each predictor independently. For the considered example, one iteration of the procedure in [U.S. Patent 5236117, col. 4, line 30-col. 9, line 25] leads

to a nearly optimal partitioning of the different predictors as follows:

$$SL_{opt}^{-1} = [l_1, l_2],\ \overline{SL_{opt}^{-1}} = [l_3, l_4],$$

$$SL_{opt}^{0} = [l_1, l_2, l_3],\ \overline{SL_{opt}^{0}} = [l_4],\ SL_{opt}^{+1} =$$

$$[l_1, l_2, \text{ and } \overline{SL_{opt}^{+1}} = [l_3, l_4].$$

**[0054]** Now, for each one of the predictors $l^k$, the training data at the current node may be split into two parts based on the partitioning $SL_{opt}^{k}, SL_{opt}^{k}$ and the probability of these two child nodes is given as:

$$Pr(SL_{opt}^{k}) = \sum_{k=1}^{3} \sum_{j \in SL_{opt}^{k}} Pr(l^k = l_j, p = p_k),$$

and

$$Pr(\overline{SL_{opt}^{k}}) \sum_{k=1}^{3} \sum_{j\ l_j \in SL_{opt}^{k}} Pr(l^k = l_j, p = p_k).$$

**[0055]** Further, the class distribution conditioned on the partitioning, at the two child nodes may be calculated as follows:

$$Pr(p = p_m / SL_{opt}^{k}) = \sum_{l_j \in SL_{opt}^{k}} Pr(l^k = l_j, p = p_m) / Pr(SL_{opt}^{k})$$

and

$$Pr(p = p_m / \overline{SL_{opt}^{k}}) = \sum_{l_j \in SL_{opt}^{k}} Pr(l^k = l_j, p = p_m) / Pr(\overline{SL_{opt}^{k}})$$

**[0056]** The entropy for each of these child nodes can be calculated just as for the parent node and the average entropy of the two child nodes computed as $H_{avg}^{k} = Pr(SL_{opt}^{k})\ H(p/SL_{opt}^{k}) + Pr(\overline{SL_{opt}^{k}})\ H(p/\overline{SL_{opt}^{k}})$. For the considered example, these quantities are tabulated in Table 4 below.

| | $p_1$ | $p_2$ | $p_3$ |
|---|---|---|---|
| $Pr(p/SL_{opt}^{-1})$ | 0.358 | 0.5 | 0.142 |
| $Pr(p/SL_{opt}^{0})$ | 0.312 | 0.188 | 0.5 |
| $Pr(p/SL_{opt}^{0})$ | 0.418 | 0.396 | 0.187 |
| $Pr(p/SL_{opt}^{0})$ | 0.117 | 0.177 | 0.707 |

(continued)

| | $p_1$ | $p_2$ | $p_3$ |
|---|---|---|---|
| $Pr(p/SL^{+1}_{opt})$ | 0.394 | 0.465 | 0.141 |
| $Pr(p/SL^{+1}_{opt})$ | 0.28 | 0.22 | 0.5 |

$$Pr(\ SL^{-1}_{opt}\ ) = 0.467\ Pr(\ \overline{SL^{-1}_{opt}}\ ) = 0.533)$$

$$H(p/\ SL^{-1}_{opt}\ ) = 1.43\ H(p/\ \overline{SL^{-1}_{opt}}\ ) = 1.477\ H^{-1}_{avg} = 1.455$$

$$Pr(\ SL^{0}_{opt}\ ) = 1.717\ Pr(\ \overline{SL^{0}_{opt}}\ ) = 0.283$$

$$H(p/\ SL^{0}_{opt}\ ) = 1.508\ H(p/\ \overline{SL^{0}_{opt}}\ ) = 1.158\ H^{0}_{avg} = 1.409$$

$$Pr(\ SL^{+1}_{opt}\ ) - 0.467\ PR(\ \overline{SL^{+1}_{opt}} - 0.533$$

$$H(p/\ SL^{+1}_{opt}\ ) - 1.442\ H(p/\ \overline{SL^{+1}_{opt}}\ ) - 1.495\ H^{+1}_{avg} - 1.470$$

**[0057]** In Block 4, the reduction in class uncertainty associated with the best question for each predictor is tabulated, and the predictor which provides the largest reduction in uncertainty is selected. The reduction in uncertainty due to a partitioning based on $SL^{k}_{avg}$ is computed as $H(p\text{-}H^{k}_{avg}$. For the considered example, we have H(p) = 1.58, $H^{-1}_{avg}$ = 1.455, $H^{0}_{avg}$ = 1.409 and $H^{+1}_{avg}$ = 1.470. Hence, the selected predictor is $1^0$, as this gives the maximum reduction in the uncertainty of the predicted class.

**[0058]** In Block 5, the training data at the current node is partitioned into two parts on the basis of the optimal partitioning of the selected predictor at the current node.

**[0059]** Subsequently, depending on the class uncertainty and the amount of training data at a child node, the process goes back to Block 2, and starts again by recomputing the joint distribution on the basis of only the training data at the child node. The processing at a child node terminates when the class uncertainty at the child node falls below a specified threshold, or if the amount of training data at a child node falls below a specified threshold.

**[0060]** Fig. 5 schematically shows an apparatus for constructing the decision tree. The apparatus may comprise of, for example, an appropriately programmed computer system. In this example, the apparatus comprises of a general purpose digital processor 8 having a data entry keyboard 9, a display 10, a random access memory 11, and a storage device 12. From the training data, processor 8 computes the joint distribution of the predictor $l^k$ and the class value p, for the first decision tree, for all 2N+1 predictors, using all of the training data, and stores the estimated joint distribution, along with the class distribution, in storage device 12.

**[0061]** Next processor 8 computes the best partitioning of each of the predictor values such that the maximum reduction in class uncertainty is obtained due to the partitioning, according to the algorithm of [U.S. Patent 5236117]. Then processor 8 chooses the best predictor, $l^*$, and partitions the training data into two child nodes based on the best partitioning for the predictor $l^*$.

**[0062]** Still under the control of the program, the processor 10 repeats the above procedure for the data at each of the two child nodes, till the class entropy at the node falls below a specified threshold, or till the amount of training data at a node falls below a specified threshold.

**[0063]** After the decision tree is grown, still under control of the program, the processor computes a distribution on class values for every node of the decision tree, and stores it in storage device 12. The above process is then repeated to construct the second decision tree. For the case of the second decision tree, the probability distribution over all possible ranks, which is stored at every node of the tree is converted into a single number, the worst case rank of the correct phone, by choosing the worst case rank as the class value at which the cumulative probability distribution of the classes exceeds a specified threshold.

**[0064]** Fig. 6 is a block diagram of an automatic speech recognition system which utilizes the decision tree according to the present invention. The system in Fig. 6 includes a microphone 13 for converting an utterance into an electrical signal. The signal from the microphone is processed by an acoustic processor and label match 14 which finds the best-

matched acoustic label prototype from the acoustic label prototype store 15. A probability distribution on phone boundaries 16a is then produced for every time frame using the first decision tree 17a described in the invention. These probabilities are compared to a threshold and some time frames are identified as boundaries between phones. Subsequently, an acoustic score is computed 16b, for all phones between every given pair of hypothesized boundaries, and the phones are ranked on the basis of this score. Note that this score may be computed in any fashion, with the only constraint being that the score is computed using the same technique as was used when constructing the second decision tree. Subsequently, the second decision tree 17b is traversed for every time frame to obtain the worst case rank of the correct phone at that time, and using the phone score and phone rank computed in 16b, a shortlist of allowed phones 16c is made up for every time frame. This information is used to select a subset of acoustic word models in store 19, and a fast acoustic word match processor 18 matches the label string from the acoustic processor 14 against this subset of abridged acoustic word models to produce an output signal.

**[0065]** The output of the fast acoustic word match processor comprises of at least one word. In general, however, the fast acoustic word match processor will output a number of candidate words.

**[0066]** Each word produced by the fast acoustic word match processor 18 is input into a word context match 20 which compares the word context to language models in store 21 and outputs at least one candidate word. From the recognition candidates produced by the fast acoustic match and the language model, the detailed acoustic match 22 matches the label string from the acoustic processor 14 against detailed acoustic word models in store 23 and outputs a word string corresponding to an utterance.

**[0067]** Fig. 7 describes Blocks 16a-c and 17a-b in further detail. Given the acoustic label string from the acoustic processor 14, the context-dependent boundary estimation process 16 traverses the first decision tree 17a for every time frame using the label at the current time and the labels at the adjacent times as the predictors, until it reaches a terminal node of the tree. Then the probability that the current time is a phone boundary is picked up from the stored class distribution at the leaf, and compared to a threshold. If the probability is larger than the threshold, it is hypothesized that the current time is a phone boundary.

**[0068]** Subsequently, an acoustic score is computed for every phone between every pair of boundary points and the phones are ranked on the basis of these scores. One of several techniques could be used to compute this score, of example, the usual markov based computation could be used, or a channel-bank-based computation as described in ["Channel-Bank-Based Thresholding to Improve Search Time in the Fast Match", IBM TDB pp. 113-114, vol. 37, No. 02A, Feb. 1994] could be used, or a decision-tree-based scoring mechanism, as described in [D. Nahamoo, M. Padmanabhan, M. A. Picheny, P. S. Gopalkrishnan, "A Decision Tree Based Pruning Strategy for the Acoustic Fast Match", IBM Attorney Docket No. YO 996-059]; the only constraint on the scoring mechanism is that the same mechanism should be used as was used when obtaining the training records for the second decision tree.

**[0069]** Subsequently, the second decision tree 17b is traversed for every time frame, using the label at the current time and at the adjacent times as the predictors, till a terminal node of the tree is reached. The worst case rank of the correct phone is read from the data stored at this node and taken to be the worst case rank of the correct phone at this time. Subsequently, the worst of the worst-case ranks between any two adjacent hypothesized phone boundaries is taken to be the worst case rank of the correct phone in the segment between the phone boundaries. All the phones whose ranks are worse than this worst case rank are then discarded in the current segment, and a shortlist of allowed phones is made up for the segment.

**[0070]** Now, it is often the case that some phones are very similar and may be easily confused with each other. Lists of such confusable phones can be made from the training data, and the shortlist described above may be augmented by adding in these lists of confusable phones. For instance, if the rank of any one element in a list of confusable phones is better than the worst case rank, the entire set of confusable phones are included in the short list.

**[0071]** It should be understood that the foregoing description is only illustrative of the invention. Various alternatives and modifications can be devised by those skilled in the art without departing from the invention. Accordingly, the present invention is intended to embrace all such alternatives, modifications and variances which fall within the scope of the appended claims.

Annex

**[0072]**

TABLE 1

THE TWO LETTERS ROUGHLY REPRESENT THE SOUND OF THE ELEMENT.
TWO DIGITS ARE ASSOCIATED WITH VOWELS:
FIRST: STRESS OF SOUND
SECOND: CURRENT IDENTIFICATION NUMBER
ONE DIGIT ONLY IS ASSOCIATED
WITH CONSONANTS
SINGLE DIGIT: CURRENT IDENTIFICATION NUMBER

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 001 | AA11 | 029 | BX2- | 057 | EH02 | 148 | TX5- | 176 | XX11 |
| 002 | AA12 | 030 | BX3- | 058 | EH11 | 149 | TX6- | 177 | XX12 |
| 003 | AA13 | 031 | BX4- | 059 | EH12 | 150 | UH01 | 178 | XX13 |
| 004 | AA14 | 032 | BX5- | 060 | EH13 | 151 | UH02 | 179 | XX14 |
| 005 | AA15 | 033 | BX6- | 061 | EH14 | 152 | UH11 | 180 | XX15 |
| 006 | AE11 | 034 | BX7- | 062 | EH15 | 153 | UH12 | 181 | XX16 |
| 007 | AE12 | 035 | BX8- | 126 | RX1- | 154 | UH13 | 182 | XX17 |
| 008 | AE13 | 036 | BX9- | 127 | SH1- | 155 | UH14 | 183 | XX1 |
| 009 | AE14 | 037 | DH1- | 128 | SH2- | 156 | UU11 | 184 | XX19 |
| 010 | AE15 | 038 | DH2- | 129 | SX1- | 157 | UU12 | 185 | XX2- |
| 011 | AW11 | 039 | DQ1- | 130 | SX2- | 158 | UXG1 | 186 | XX20 |
| 012 | AW12 | 040 | DQ2- | 131 | SX3- | 159 | UXG2 | 187 | XX21 |
| 013 | AW13 | 041 | DQ3- | 132 | SX4- | 160 | UX11 | 186 | XX22 |
| 014 | AX11 | 042 | DQ4- | 133 | SX5- | 161 | UX12 | 189 | XX23 |
| 015 | AX12 | 043 | DX1- | 134 | SX6- | 162 | UX13 | 190 | XX24 |
| 016 | AX13 | 044 | DX2- | 135 | SX7- | 163 | VX1- | 191 | XX3- |
| 017 | AX14 | 045 | EE01 | 136 | TH1- | 164 | VX2- | 192 | XX4- |
| 018 | AX15 | 046 | EE02 | 137 | TH2- | 165 | VX3- | 193 | XX5- |
| 019 | AX16 | 047 | EE11 | 138 | TH3- | 166 | VX4- | 194 | XX6- |
| 020 | AX17 | 048 | EE12 | 139 | TH4- | 167 | WX1- | 195 | XX7- |
| 021 | BQ1- | 049 | EE13 | 140 | TH5- | 168 | WX2- | 196 | XX8- |
| 022 | BQ2- | 050 | EE14 | 141 | TQ1- | 169 | WX3- | 197 | XX9- |
| 023 | BQ3- | 051 | EE15 | 142 | TQ2- | 170 | WX4- | 198 | ZX1- |
| 024 | BQ4- | 052 | EE16 | 143 | TX3- | 171 | WX5- | 199 | ZX2- |
| 025 | BX1- | 053 | EE17 | 144 | TX1- | 172 | WX6- | 200 | ZX3- |
| 026 | BX10 | 054 | EE18 | 145 | TX2- | 173 | WX7- | | |
| 027 | BX11 | 055 | EE19 | 146 | TX3- | 174 | XX1- | | |
| 028 | BX12 | 056 | EH01 | 147 | TX4- | 175 | XX10 | | |

## Claims

1. A method of recognizing speech, comprising the steps of:

a) inputting a plurality of words of training data;

b) training one or more binary first decision trees to ask a maximally informative question at each node based upon contextual information in the training data, wherein each binary first decision tree may correspond to a different time in a sequence of the training data;

c) traversing one of the decision trees for every time frame of an input sequence of speech to determine a probability distribution for every time frame, the probability distribution being the probability that a node is a

phone boundary;

d) comparing the probabilities associated with the time frames with a threshold for identifying some time frames as boundaries between phones;

e) providing an acoustic score for all phones between every given pair of boundaries;

f) ranking the phones on the basis of this score;

g) outputting a recognition result in response to the score.

2. The method of claim 1 further including the steps of:

h) traversing one or more of a second set of decision trees for every time frame on an input sequence of speech to determine a second probability distribution, the probability distribution being a distribution over all possible ranks that the correct phone can take for obtaining a worst case rank of a correctly recognized phone by choosing the worst case rank as the class value at which the cumulative probability distribution of the classes exceeds a specified threshold;

i) assigning as the absolute worst case rank of the worst case ranks between any two adjacent phone boundaries the worst case rank of the correctly recognized phone between the phone boundaries;

j) discarding all phones whose rank is worse than this absolute worst case rank in the current segment;

k) making a short list of phones for the segment;

l) outputting a recognition result in response to the short list of the recognition result being a short list of words.

3. The method of claim 1, further including the steps of:

h) traversing one or more of a second set of decision trees for every time frame on an input sequence of speech to determine a second probability distribution, the probability distribution being a distribution over all possible ranks that a phone can take for obtaining a worst case rank of a correctly recognized phone by choosing the worst case rank as the class value at which the cumulative probability distribution of the classes exceeds a specified threshold;

i) assigning as the absolute worst case rank of the worst case ranks between any two adjacent phone boundaries the worst case rank of the correctly recognized phone between the phone boundaries;

j) discarding all phone boundaries whose rank is worse than this absolute worst case rank in the current segment;

k) making a short list of phones for the segment;

l) comparing constituent phones of a word in a vocabulary to see if the word lies in the short list and making up a short list of words;

l) outputting a recognition result by comparing the words of the short list with a language model to determine the most probable word match for the input sequence of speech.

4. A method for recognizing speech, comprising the steps of:

a) entering a string of utterances constituting training data;

b) converting the utterances of the training data to electrical signals;

c) representing the electrical signal of the training data as prototype quantized feature vectors, one feature vector representing a given time frame;

d) assigning to each prototype feature vector a class label associated with the prototype quantized feature vector;

e) forming one or more binary decision trees for different times in the training data, each tree having a root node and a plurality of child nodes, comprising the steps of:

i. creating a set of training records comprising 2K+1 predictors, $l^k$, and one predicted class, p, where the 2K+1 predictors are feature vector labels at 2K+1 consecutive times t-K, ..., t, ..., t, ..., t+K, and the predicted class is a binary record indicator whether time t is associated with a phone boundary in the case of the first decision tree or is associated with the correct phone in the case of the second decision tree;

ii. computing the estimated joint distribution of predictors $l^k$ and phone p for 2K+1 predictors using the training data, wherein the predictors are feature vector labels at times t-K, ..., t, ..., t+K and p is the phone at time t;

iii. storing the estimated joint distribution of $l^k$ and p and a corresponding distribution for each predictor $l^k$ at the root node;

iv. computing the best partitioning of the values that predictor $l^k$ can take for each $l^k$ to minimize phone uncertainty at each node;

v. choosing the predictor $l^k$ whose partitioning results in the lowest uncertainty and partitioning the training data into two child nodes based on the computed-based partitioning $l^k$, each child node being assigned a class distribution based on the training data at the child node;

f) repeating for each child node if the amount of training data at the child node is greater than a threshold;

g) inputting an utterance to be recognized;

h) converting the utterance into an electrical signal;

i) representing the electrical signal as a series of quantized feature vectors;

j) matching the series of quantized feature vectors against the stored prototype feature vectors to determine a closest match and assigning an input label to each of the series of feature vectors corresponding to the label of the closest matching prototype feature vector;

k) traversing one of the decision trees for every time frame of an input sequence of speech to determine a probability distribution for every time frame, the probability distribution being the probability that a node is a phone boundary;

l) comparing the probabilities associated with the time frames with a threshold for identifying some time frames as boundaries between phones;

m) providing an acoustic score for all phones between every given pair of boundaries;

n) ranking the phones on the basis of this score;

o) outputting a recognition result in response to the score.

5. The method of claim 4 further including the steps of:

traversing one or more of a second set of decision trees for every time frame on an input sequence of speech to determine a second probability distribution, the probability distribution being a distribution over all possible ranks that a phone can take for obtaining a worst case rank of a correctly recognized phone by choosing the worst case rank as the class value at which the cumulative probability distribution of the classes exceeds a specified threshold;

assigning as the absolute worst case rank of the worst case ranks between any two adjacent phone boundaries the worst case rank of the correctly recognized phone between the phone boundaries;

discarding all phone boundaries whose rank is worse than this absolute worst case rank in the current segment;

making a short list for the segment;

outputting a recognition result in response to the short list.

6. An apparatus for recognizing speech, comprising:

a) means for inputting a plurality of words of training data;

b) means for training one or more binary first decision trees to ask a maximally informative question at each node based upon contextual information in the training data, wherein each binary first decision tree may correspond to a different time in a sequence of the training data;

c) means for traversing one of the decision trees for every time frame of an input sequence of speech to determine a probability distribution for every time frame, the probability distribution being the probability that a node is a phone boundary;

d) means for comparing the probabilities associated with the time frames with a threshold for identifying some time frames as boundaries between phones;

e) means for providing an acoustic score for all phones between every given pair of boundaries;

f) means for ranking the phones on the basis of this score;

g) means for outputting a recognition result in response to the score.

7. The apparatus of claim 6 further including:

h) means for traversing one or more of a second set of decision trees for every time frame on an input sequence of speech to determine a second probability distribution, the probability distribution being a distribution over all possible ranks that the correct phone can take for obtaining a worst case rank of a correctly recognized phone by choosing the worst case rank as the class value at which the cumulative probability distribution of the classes exceeds a specified threshold;

i) means for assigning as the absolute worst case rank of the worst case ranks between any two adjacent phone boundaries the worst case rank of the correctly recognized phone between the phone boundaries;

j) means for discarding all phones whose rank is worse than this absolute worst case rank in the current segment;

k) means for making a short list of phones for the segment;

l) means for outputting a recognition result in response to the short list of the recognition result being a short list of words.

8. The apparatus of claim 6, further including:

h) means for traversing one or more of a second set of decision trees for every time frame on an input sequence of speech to determine a second probability distribution, the probability distribution being a distribution over all possible ranks that a phone can take for obtaining a worst case rank of a correctly recognized phone by choosing the worst case rank as the class value at which the cumulative probability distribution of the classes exceeds a specified threshold;

i) means for assigning as the absolute worst case rank of the worst case ranks between any two adjacent

phone boundaries the worst case rank of the correctly recognized phone between the phone boundaries;

j) means for discarding all phone boundaries whose rank is worse than this absolute worst case rank in the current segment;

k) means for making a short list of phones for the segment;

l) means for comparing constituent phones of a word in a vocabulary to see if the word lies in the short list and making up a short list of words;

m) means for outputting a recognition result by comparing the words of the short list with a language model to determine the most probable word match for the input sequence of speech.

**9.** An apparatus for recognizing speech, comprising:

a) means for entering a string of utterances constituting training data;

b) means for converting the utterances of the training data to electrical signals;

c) means for representing the electrical signal of the training data as prototype quantized feature vectors, one feature vector representing a given time frame;

d) means for assigning to each prototype feature vector a class label associated with the prototype quantized feature vector;

e) means for forming one or more binary decision trees for different times in the training data, each tree having a root node and a plurality of child nodes, comprising the steps of:

i. means for creating a set of training records comprising 2K+1 predictors, $l^k$, and one predicted class, p, where the 2K+1 predictors are feature vector labels at 2K+1 consecutive times t-K, ..., t, ..., t, ..., t+K, and the predicted class is a binary record indicator whether time t is associated with a phone boundary in the case of the first decision tree or is associated with the correct phone in the case of the second decision tree;

ii. means for computing the estimated joint distribution of predictors $l^k$ and phone p for 2K+1 predictors using the training data, wherein the predictors are feature vector labels at times t-K, ..., t, ..., t+K and p is the phone at time t;

iii. means for storing the estimated joint distribution of $l^k$ and p and a corresponding distribution for each predictor $l^k$ at the root node;

iv. means for computing the best partitioning of the values that predictor $l^k$ can take for each $l^k$ to minimize phone uncertainty at each node;

v. means for choosing the predictor $l^k$ whose partitioning results in the lowest uncertainty and partitioning the training data into two child nodes based on the computed-based partitioning $l^k$, each child node being assigned a class distribution based on the training data at the child node;

f) means for repeating for each child node if the amount of training data at the child node is greater than a threshold;

g) means for inputting an utterance to be recognized;

h) means for converting the utterance into an electrical signal;

i) means for representing the electrical signal as a series of quantized feature vectors;

j) means for matching the series of quantized feature vectors against the stored prototype feature vectors to determine a closest match and assigning an input label to each of the series of feature vectors corresponding

to the label of the closest matching prototype feature vector;

k) means for traversing one of the decision trees for every time frame of an input sequence of speech to determine a probability distribution for every time frame, the probability distribution being the probability that a node is a phone boundary;

l) means for comparing the probabilities associated with the time frames with a threshold for identifying some time frames as boundaries between phones;

m) means for providing an acoustic score for all phones between every given pair of boundaries;

n) means for ranking the phones on the basis of this score;

o) means for outputting a recognition result in response to the score.

**10.** The apparatus of claim 9 further including:
means for traversing one or more of a second set of decision trees for every time frame on an input sequence of speech to determine a second probability distribution, the probability distribution being a distribution over all possible ranks that a phone can take for obtaining a worst case rank of a correctly recognized phone by choosing the worst case rank as the class value at which the cumulative probability distribution of the classes exceeds a specified threshold;
means for assigning as the absolute worst case rank of the worst case ranks between any two adjacent phone boundaries the worst case rank of the correctly recognized phone between the phone boundaries;
means for discarding all phone boundaries whose rank is worse than this absolute worst case rank in the current segment;
means for making a short list for the segment;
means for outputting a recognition result in response to the short list.

## Patentansprüche

**1.** Ein Verfahren zur Spracherkennung, das folgende Schritte umfaßt:

a) Eingabe mehrerer Wörter der Trainingsdaten;

b) Training eines oder mehrerer binärer erster Entscheidungsbäume, um an jedem Knoten auf der Grundlage von Kontextdaten innerhalb der Trainingsdaten eine möglichst informative Frage zu stellen, wobei jeder binäre erste Entscheidungsbaum einem anderen Zeitpunkt in einer Sequenz der Trainingsdaten entsprechen kann;

c) Durchlaufen eines Entscheidungsbaums für jeden Zeitrahmen einer Spracheingabesequenz, um für jeden Zeitrahmen eine Wahrscheinlichkeitsverteilung zu bestimmen, wobei die Wahrscheinlichkeitsverteilung die Wahrscheinlichkeit ist, daß ein Knoten eine Phonemgrenze ist;

d) Vergleich der Wahrscheinlichkeiten der Zeitrahmen mit einem Schwellenwert zur Bestimmung einiger Zeitrahmen als Grenzen zwischen Phonemen;

e) Bereitstellung einer akustischen Trefferzahl für alle Phoneme zwischen jedem gegebenen Grenzenpaar

f) Klassifizierung der Phoneme auf der Grundlage dieser Trefferzahl;

g) Ausgabe eines Erkennungsergebnisses in Abhängigkeit dieser Trefferzahl.

**2.** Das Verfahren gemäß Anspruch 1, das weiterhin folgende Schritte umfaßt:

h) Durchlaufen eines Entscheidungsbaums oder mehrerer Entscheidungsbäume aus einer zweiten Gruppe von Entscheidungsbäumen für jeden Zeitrahmen in einer Spracheingabesequenz zur Bestimmung einer zweiten Wahrscheinlichkeitsverteilung, wobei die Wahrscheinlichkeitsverteilung eine Verteilung über alle Klassen ist, die für das korrekte Phonem möglich sind, um eine Klasse des schlimmsten Falls eines richtig erkannten

Phonems einzuholen, indem die Klasse des schlimmsten Falls als Klassenwert gewählt wird, bei dem die kumulative Wahrscheinlichkeitsverteilung der Klassen einen bestimmten Schwellenwert überschreitet;

i) Unter den Klassen des schlimmsten Falls Bestimmung zur Klasse des absolut schlimmsten Falls zwischen zwei beliebigen nebeneinander liegenden Phonemgrenzen der Klasse des schlimmsten Falls des richtig erkannten Phonems zwischen den Phonemgrenzen;

j) Aussparung aller Phoneme, deren Klasse schlimmer ist als diese Klasse des absolut schlimmsten Falls im aktuellen Segment;

k) Erstellung einer Kurzliste von Phonemen für das Segment;

l) Ausgabe eines Erkennungsergebnisses, wenn die Kurzliste des Erkennungsergebnisses eine Kurzliste aus Wörtern ist.

3. Verfahren gemäß Anspruch 1, das weiterhin die folgenden Schritte umfaßt:

h) Durchlaufen eines oder mehrerer Entscheidungsbäume aus einer zweiten Gruppe an Entscheidungsbäumen für jeden Zeitrahmen einer Spracheingangssequenz zur Bestimmung einer zweiten Wahrscheinlichkeitsverteilung, wobei die Wahrscheinlichkeitsverteilung eine Verteilung über alle möglichen Klassen ist, in die ein Phonem aufgenommen werden kann, um eine Klasse des schlimmsten Falls eines richtig erkannten Phonems zu erhalten, und zwar durch Bestimmung der Klasse des schlimmsten Falls zum Klassenwert, bei dem die kumulative Wahrscheinlichkeitsverteilung der Klassen einen bestimmten Schwellenwert überschreitet;

i) Unter den Klassen des schlimmsten Falls Bestimmung zur Klasse des absolut schlimmsten Falls zwischen zwei beliebigen nebeneinander liegenden Phonemgrenzen der Klasse des schlimmsten Falls des richtig erkannten Phonems zwischen den Phonemgrenzen;

j) Aussparung aller Phoneme, deren Klasse schlimmer ist als diese Klasse des absolut schlimmsten Falls im aktuellen Segment;

k) Erstellung einer Kurzliste von Phonemen für das Segment;

l) Vergleich bestandteilbildender Phoneme eines Wortes in einem Vokabular, um festzustellen, ob das Wort in der Kurzliste enthalten ist, und Erstellung einer Kurzliste von Wörtern;

l) Ausgabe eines Erkennungsergebnisses durch Vergleich der Wörter aus der Kurzliste mit einem Sprachmodell, um die am meisten wahrscheinliche Wortübereinstimmung für die Spracheingangssequenz zu bestimmen.

4. Ein Verfahren zur Spracherkennung, das die folgenden Schritte umfaßt:

a) Eingabe eines Strings von Sprachelementen, die Trainingsdaten darstellen;

b) Umwandlung der Elemente der Trainingsdaten in elektrische Signale;

c) Darstellung des elektrischen Signals der Trainingsdaten als prototyp-quantisierte Eigenschaftsvektoren, wobei ein Eigenschaftsvektor einen gegebenen Zeitrahmen darstellt;

d) Zuordnung eines Klassenlabels für den prototypquantisierten Eigenschaftsvektor zu jedem Prototyp-Eigenschaftsvektor;

e) Aufbau eines oder mehrerer Entscheidungsbäume für unterschiedliche Zeiten in den Trainingsdaten, wobei jeder Baum einen Wurzelknoten und eine Mehrzahl an Kindknoten aufweist, bestehend aus den folgenden Schritten:

i. Bildung einer Gruppe von Trainingsaufzeichnungen, die 2K+1 Prädiktoren, $1^k$, und eine vorausgesagte Klasse, p, umfassen, wobei die 2K+1 Prädiktoren Eigenschaftsvektorlabels an 2K+1 aufeinanderfolgen-

den Zeiten t-K, ..., t, ..., t+K sind und die vorausgesagte Klasse eine binäre Aufzeichnungsanzeige darüber ist, ob der Zeitpunkt t zu einer Phonemgrenze im Fall des ersten Entscheidungsbaums gehört oder zum korrekten Phonem im Fall des zweiten Entscheidungsbaums gehört;

ii. Berechnung der geschätzten verbundenen Verteilung der Prädiktoren $1^k$ und des Phonems p für 2K+1 Prädiktoren unter Verwendung der Trainingsdaten, wobei die Prädiktoren Eigenschaftsvektorlabels zu den Zeitpunkten t-K, ..., t, ..., t+K sind und p das Phonem zum Zeitpunkt t ist;

iii. Speicherung der geschätzten verbundenen Verteilung von $1^k$ und p und einer entsprechenden Verteilung für jeden Prädiktor $1^k$ am Wurzelknoten;

iv. Berechnung der besten Partitionierung der Werte, die der Prädiktor $1^k$ für jedes $1^k$ annehmen kann, um die Phonemungewißheit an jedem Knoten auf ein Mindestmaß zu beschränken;

v. Auswahl des Prädiktors $1^k$, dessen Partitionierung zur niedrigsten Ungewißheit führt, und Partitionierung der Trainingsdaten in zwei Kindknoten, und zwar auf der Grundlage der computergesteuerten Partitionierung $1^k$, wobei jedem Kindknoten auf der Grundlage der Trainingsdaten am Kindknoten eine Klassenverteilung zugeordnet wird;

f) Wiederholung der Bestimmung für jeden Kindknoten, ob der Umfang an Trainingsdaten am Kindknoten größer ist als ein Schwellenwert;

g) Eingabe eines Sprachelements, das erkannt werden soll;

h) Umwandlung eines Sprachelements in ein elektrisches Signal;

i) Darstellung des elektrischen Signals als Serie quantisierter Eigenschaftsvektoren;

j) Vergleich der Serie quantisierter Eigenschaftsvektoren mit den gespeicherten Prototyp-Eigenschaftsvektoren zur Bestimmung einer engsten Übereinstimmung und Zuordnung eines Eingangslabels zu jedem Vektor aus der Serie der Eigenschaftsvektoren entsprechend dem Label des am engsten übereinstimmenden Eigenschaftsvektors;

k) Durchlaufen eines Entscheidungsbaums für jeden Zeitrahmen einer Spracheingabesequenz, um für jeden Zeitrahmen eine Wahrscheinlichkeitsverteilung zu bestimmen, wobei die Wahrscheinlichkeitsverteilung die Wahrscheinlichkeit ist, daß ein Knoten eine Phonemgrenze ist;

l) Vergleich der Wahrscheinlichkeiten der Zeitrahmen mit einem Schwellenwert zur Bestimmung einiger Zeitrahmen als Grenzen zwischen Phonemen;

m) Bereitstellung einer akustischen Trefferzahl für alle Phoneme zwischen jedem gegebenen Grenzenpaar;

n) Klassifizierung der Phoneme auf der Grundlage dieser Trefferzahl;

o) Ausgabe eines Erkennungsergebnisses in Abhängigkeit dieser Trefferzahl.

**5.** Das Verfahren gemäß Anspruch 4, das weiterhin folgende Schritte umfaßt:

Durchlaufen eines Entscheidungsbaums oder mehrerer Entscheidungsbäume aus einer zweiten Gruppe von Entscheidungsbäumen für jeden Zeitrahmen in einer Spracheingabesequenz zur Bestimmung einer zweiten Wahrscheinlichkeitsverteilung, wobei die Wahrscheinlichkeitsverteilung eine Verteilung über alle Klassen ist, die für das korrekte Phonem möglich sind, um eine Klasse des schlimmsten Falls eines richtig erkannten Phonems einzuholen, indem die Klasse des schlimmsten Falls als Klassenwert gewählt wird, bei dem die kumulative Wahrscheinlichkeitsverteilung der Klassen einen bestimmten Schwellenwert überschreitet;

Unter den Klassen des schlimmsten Falls Bestimmung zur Klasse des absolut schlimmsten Falls zwischen zwei beliebigen nebeneinander liegenden Phonemgrenzen der Klasse des schlimmsten Falls des richtig erkannten Phonems zwischen den Phonemgrenzen;

Aussparung aller Phonemgrenzen, deren Klasse schlimmer ist als diese Klasse des absolut schlimmsten Falls im aktuellen Segment;

Erstellung einer Kurzliste für das Segment;

Ausgabe eines Erkennungsergebnisses als Antwort auf die Kurzliste.

6. Eine Vorrichtung zur Spracherkennung, die folgendes umfaßt:

a) Mittel zur Eingabe mehrerer Trainingsdatenwörter;

b) Mittel für das Training eines oder mehrerer binärer erster Entscheidungsbäume, um an jedem Knoten auf der Grundlage von Kontextdaten innerhalb der Trainingsdaten eine möglichst informative Frage zu stellen, wobei jeder binäre erste Entscheidungsbaum einem anderen Zeitpunkt in einer Sequenz der Trainingsdaten entsprechen kann;

c) Mittel für das Durchlaufen eines Entscheidungsbaums für jeden Zeitrahmen einer Spracheingabesequenz, um für jeden Zeitrahmen eine Wahrscheinlichkeitsverteilung zu bestimmen, wobei die Wahrscheinlichkeitsverteilung die Wahrscheinlichkeit ist, daß ein Knoten eine Phonemgrenze ist;

d) Mittel für den Vergleich der Wahrscheinlichkeiten der Zeitrahmen mit einem Schwellenwert zur Bestimmung einiger Zeitrahmen als Grenzen zwischen Phonemen;

e) Mittel für die Bereitstellung einer akustischen Trefferzahl für alle Phoneme zwischen jedem gegebenen Grenzenpaar;

f) Mittel für die Klassifizierung der Phoneme auf der Grundlage dieser Trefferzahl;

g) Mittel für die Ausgabe eines Erkennungsergebnisses in Abhängigkeit dieser Trefferzahl.

7. Die Vorrichtung gemäß Anspruch 6, die weiterhin folgendes umfaßt:

h) Mittel für das Durchlaufen eines Entscheidungsbaums oder mehrerer Entscheidungsbäume aus einer zweiten Gruppe von Entscheidungsbäumen für jeden Zeitrahmen in einer Spracheingabesequenz zur Bestimmung einer zweiten Wahrscheinlichkeitsverteilung, wobei die Wahrscheinlichkeitsverteilung eine Verteilung über alle Klassen ist, die für das korrekte Phonem möglich sind, um eine Klasse des schlimmsten Falls eines richtig erkannten Phonems einzuholen, indem die Klasse des schlimmsten Falls als Klassenwert gewählt wird, bei dem die kumulative Wahrscheinlichkeitsverteilung der Klassen einen bestimmten Schwellenwert überschreitet;

i) Unter den Klassen des schlimmsten Falls ein Mittel zur Bestimmung zur Klasse des absolut schlimmsten Falls zwischen zwei beliebigen nebeneinander liegenden Phonemgrenzen der Klasse des schlimmsten Falls des richtig erkannten Phonems zwischen den Phonemgrenzen;

j) Mittel zur Aussparung aller Phonemgrenzen, deren Klasse schlimmer ist als diese Klasse des absolut schlimmsten Falls im aktuellen Segment;

k) Mittel zur Erstellung einer Kurzliste für das Segment;

l) Mittel zur Ausgabe eines Erkennungsergebnisses, wenn die Kurzliste des Erkennungsergebnisses eine Kurzliste von Wörtern ist.

8. Die Vorrichtung gemäß Anspruch 6, die weiterhin folgendes umfaßt:

h) Mittel für das Durchlaufen eines Entscheidungsbaums oder mehrerer Entscheidungsbäume aus einer zweiten Gruppe von Entscheidungsbäumen für jeden Zeitrahmen in einer Spracheingabesequenz zur Bestimmung einer zweiten Wahrscheinlichkeitsverteilung, wobei die Wahrscheinlichkeitsverteilung eine Verteilung über alle Klassen ist, die für das korrekte Phonem möglich sind, um eine Klasse des schlimmsten Falls eines richtig

erkannten Phonems einzuholen, indem die Klasse des schlimmsten Falls als Klassenwert gewählt wird, bei dem die kumulative Wahrscheinlichkeitsverteilung der Klassen einen bestimmten Schwellenwert überschreitet;

i) Unter den Klassen des schlimmsten Falls ein Mittel zur Bestimmung zur Klasse des absolut schlimmsten Falls zwischen zwei beliebigen nebeneinander liegenden Phonemgrenzen der Klasse des schlimmsten Falls des richtig erkannten Phonems zwischen den Phonemgrenzen;

j) Mittel zur Aussparung aller Phonemgrenzen, deren Klasse schlimmer ist als diese Klasse des absolut schlimmsten Falls im aktuellen Segment;

k) Mittel zur Erstellung einer Kurzliste der Phoneme für das Segment;

l) Mittel für den Vergleich bestandteilbildender Phoneme eines Wortes in einem Vokabular, um festzustellen, ob das Wort in der Kurzliste enthalten ist, und die Erstellung einer Kurzliste von Wörtern;

l) Mittel für die Ausgabe eines Erkennungsergebnisses durch Vergleich der Wörter aus der Kurzliste mit einem Sprachmodell, um die am meisten wahrscheinliche Wortübereinstimmung für die Spracheingangssequenz zu bestimmen.

**9.** Eine Vorrichtung zur Spracherkennung, die folgendes umfaßt:

a) Mittel zur Eingabe eines Strings von Sprachelementen, die Trainingsdaten darstellen;

b) Mittel zur Umwandlung der Elemente der Trainingsdaten in elektrische Signale;

c) Mittel zur Darstellung des elektrischen Signals der Trainingsdaten als prototyp-quantisierte Eigenschaftsvektoren, wobei ein Eigenschaftsvektor einen gegebenen Zeitrahmen darstellt;

d) Mittel zur Zuordnung eines Klassenlabels für den prototyp-quantisierten Eigenschaftsvektor zu jedem Prototyp-Eigenschaftsvektor;

e) Mittel zum Aufbau eines oder mehrerer binärer Entscheidungsbäume für unterschiedliche Zeiten in den Trainingsdaten, wobei jeder Baum einen Wurzelknoten und eine Mehrzahl an Kindknoten aufweist, bestehend aus den folgenden Schritten:

i. Mittel zur Bildung einer Gruppe von Trainingsaufzeichnungen, die 2K+1 Prädiktoren, $1^k$, und eine vorausgesagte Klasse, p, umfassen, wobei die 2K+1 Prädiktoren Eigenschaftsvektorlabels an 2K+1 aufeinanderfolgenden Zeiten t-K, ..., t, ..., t+K sind und die vorausgesagte Klasse eine binäre Aufzeichnungsanzeige darüber ist, ob der Zeitpunkt t zu einer Phonemgrenze im Fall des ersten Entscheidungsbaums gehört oder zum korrekten Phonem im Fall des zweiten Entscheidungsbaums gehört;

ii. Mittel zur Berechnung der geschätzten verbundenen Verteilung der Prädiktoren $1^k$ und des Phonems p für 2K+1 Prädiktoren unter Verwendung der Trainingsdaten, wobei die Prädiktoren Eigenschaftsvektorlabels zu den Zeitpunkten t-K, ..., t, ..., t+K sind und p das Phonem zum Zeitpunkt t ist;

iii. Mittel zur Speicherung der geschätzten verbundenen Verteilung von $1^k$ und p und einer entsprechenden Verteilung für jeden Prädiktor $1^k$ am Wurzelknoten;

iv. Mittel zur Berechnung der besten Partitionierung der Werte, die der Prädiktor $1^k$ für jedes $1^k$ annehmen kann, um die Phonemungewißheit an jedem Knoten auf ein Mindestmaß zu beschränken;

v. Mittel zur Auswahl des Prädiktors $1^k$, dessen Partitionierung zur niedrigsten Ungewißheit führt, und Partitionierung der Trainingsdaten in zwei Kindknoten, und zwar auf der Grundlage der computergesteuerten Partitionierung $1^k$, wobei jedem Kindknoten auf der Grundlage der Trainingsdaten am Kindknoten eine Klassenverteilung zugeordnet wird;

f) Mittel zur Wiederholung der Bestimmung für jeden Kindknoten, ob der Umfang an Trainingsdaten am Kind-

knoten größer ist als ein Schwellenwert;

g) Mittel zur Eingabe eines Sprachelements, das erkannt werden soll;

h) Mittel zur Umwandlung eines Sprachelements in ein elektrisches Signal;

i) Mittel zur Darstellung des elektrischen Signals als Serie quantisierter Eigenschaftsvektoren;

j) Mittel zum Vergleich der Serie quantisierter Eigenschaftsvektoren mit den gespeicherten Prototyp-Eigenschaftsvektoren zur Bestimmung einer engsten Übereinstimmung und Zuordnung eines Eingangslabels zu jedem Vektor aus der Serie der Eigenschaftsvektoren entsprechend dem Label des am engsten übereinstimmenden Eigenschaftsvektors;

k) Mittel für das Durchlaufen eines Entscheidungsbaums für jeden Zeitrahmen einer Spracheingabesequenz, um für jeden Zeitrahmen eine Wahrscheinlichkeitsverteilung zu bestimmen, wobei die Wahrscheinlichkeitsverteilung die Wahrscheinlichkeit ist, daß ein Knoten eine Phonemgrenze ist;

l) Mittel für den Vergleich der Wahrscheinlichkeiten der Zeitrahmen mit einem Schwellenwert zur Bestimmung einiger Zeitrahmen als Grenzen zwischen Phonemen;

m) Mittel zur Bereitstellung einer akustischen Trefferzahl für alle Phoneme zwischen jedem gegebenen Grenzenpaar;

n) Mittel zur Klassifizierung der Phoneme auf der Grundlage dieser Trefferzahl;

o) Mittel zur Ausgabe eines Erkennungsergebnisses in Abhängigkeit dieser Trefferzahl.

**10.** Die Vorrichtung gemäß Anspruch 9, die weiterhin folgendes umfaßt:

Mittel für das Durchlaufen eines Entscheidungsbaums oder mehrerer Entscheidungsbäume aus einer zweiten Gruppe von Entscheidungsbäumen für jeden Zeitrahmen in einer Spracheingabesequenz zur Bestimmung einer zweiten Wahrscheinlichkeitsverteilung, wobei die Wahrscheinlichkeitsverteilung eine Verteilung über alle Klassen ist, die für das korrekte Phonem möglich sind, um eine Klasse des schlimmsten Falls eines richtig erkannten Phonems einzuholen, indem die Klasse des schlimmsten Falls als Klassenwert gewählt wird, bei dem die kumulative Wahrscheinlichkeitsverteilung der Klassen einen bestimmten Schwellenwert überschreitet;

Unter den Klassen des schlimmsten Falls ein Mittel zur Bestimmung zur Klasse des absolut schlimmsten Falls zwischen zwei beliebigen nebeneinander liegenden Phonemgrenzen der Klasse des schlimmsten Falls des richtig erkannten Phonems zwischen den Phonemgrenzen;

Mittel zur Aussparung aller Phonemgrenzen, deren Klasse schlimmer ist als diese Klasse des absolut schlimmsten Falls im aktuellen Segment;

Mittel zur Erstellung einer Kurzliste für das Segment;

Mittel zur Ausgabe eines Erkennungsergebnisses in Reaktion auf die Kurzliste.

**Revendications**

**1.** Méthode de reconnaissance de la parole, comportant les phases qui consistent à :

a) entrer une pluralité de mots des données de formation ;

b) former un ou plusieurs premiers arbres binaires de décision pour poser la question la plus instructive à chaque noeud, en se basant sur l'information contextuelle dans les données de formation, où chaque premier arbre binaire de décision peut correspondre à un temps différent dans une séquence des données de forma-

tion;

c) traverser un des arbres de décision pour chaque tranche de temps d'une séquence d'entrée du discours pour déterminer une distribution de probabilité pour chaque tranche de temps, la distribution de probabilité étant la probabilité qu'un noeud soit une limite de son ;

d) comparer les probabilités associées aux tranches de temps à un seuil pour identifier quelques tranche de temps comme limites de sons ;

e) fournir un résultat acoustique pour toutes les limites de temps entre chaque paire donnée de limites ;

f) classer les sons sur la base de ce résultat;

g) sortir un résultat de reconnaissance en réponse au résultat.

2. Méthode selon la revendication 1 comprenant en outre les phases qui consistent à:

h) parcourir un ou plusieurs d'un deuxième ensemble d'arbres de décision pour chaque tranche de temps sur une séquence d'entrée du discours pour déterminer une deuxième distribution de probabilité, la probabilité de distribution étant une distribution au-dessus de tous les rangs possibles que le son correct peut avoir pour obtenir un plus classement parmi les pires d'un son correctement identifié en choisissant le pire classement comme valeur type à laquelle la distribution de probabilité cumulative des classes dépasse un seuil précis ;

i) attribuer comme le pire classement possible parmi les classements les pires entre deux limites de son adjacentes quelconques, le pire classement du son correctement identifié entre les limites du son ;

j) négliger tous les sons dont le rang est plus mauvais que ce pire classement absolu dans le segment courant ;

k) faire une courte liste des sons pour le segment;

l) sortir un résultat d'identification si la liste courte du résultat d'identification étant une liste courte de mots.

3. Méthode selon la revendication 1, comprenant en outre les phases qui consistent à:

h) traverser un ou plusieurs arbres d'un deuxième ensemble d'arbres de décision pour chaque tranche de temps sur une séquence d'entrée du discours pour déterminer une deuxième distribution de probabilité, la distribution de probabilité étant une distribution sur tous les rangs possibles qu'un son peut prendre pour obtenir un pire classement d'un son correctement identifié en choisissant le pire classement comme valeur type à laquelle la distribution de probabilité cumulative des types dépasse un seuil précis ;

i) attribuer en tant que pire classement possible de tous les pires classements entre deux limites de sons adjacentes quelconques, le pire classement du son correctement reconnu entre les limites du son ;

j) négliger tous les sons dont le rang est plus mauvais que ce pire classement absolu dans le segment courant ;

k) faire une courte liste des sons pour le segment;

l) comparer les sons constituant un mot avec un vocabulaire pour voir si le mot se trouve dans la liste courte et établir une courte liste de mots ;

m) sortir un résultat de reconnaissance en comparant les mots de la liste courte avec un modèle de langue pour déterminer la concordance la plus probable d'un mot pour la séquence de parole entrée.

4. Méthode de reconnaissance de la parole, comprenant les phases qui consistent à :

a) entrer un suite d'expressions constituant des données de formation ;

b) convertir les expressions des données de formation en signaux électriques ;

c) représenter du signal électrique des données de formation par des vecteurs prototypes, un vecteur représentant une tranche de temps donnée ;

d) attribuer à chaque vecteur de caractéristique prototype, une étiquette de classe associée au vecteur prototype de caractéristique quantifiée ;

e) former un ou plusieurs arbres de décision binaires pour différents temps dans les données de formation, chaque arbre ayant un noeud racine et une pluralité de noeuds enfants, comprenant les phases suivantes :

i) créer un ensemble d'enregistrements de formation comportant 2K+1 les prédiseurs, 1k, et une classe prévue, p, où les prédiseurs 2K+1 sont des étiquettes de vecteur de dispositif 2K+1 aux temps consécutifs t-K, ..., t,..., t,..., :t+K, et la classe prévue est un indicateur d'enregistrement binaire si le temps t est associé à une limite de son dans le cas du premier arbre de décision ou est associé au son correct dans le cas du deuxième arbre de décision;

ii. calculer la distribution commune estimée des prédiseurs $1^k$ et du son p pour les prédiseurs 2K+1 en utilisant les données de formation, où les prédiseurs sont les étiquettes de vecteur de caractéristiques aux temps t-K,..., t,..., t+K et p est le son au temps t;

iii. enregistrer la distribution commune estimée de $1^k$ et de p et une distribution correspondante pour chaque prédiseur $1^k$ au noeud racine;

iv. calculer la meilleure division des valeurs que le prédiseur $1^k$ peut prendre pour chaque $1^k$ pour minimiser l'incertitude sur le son à chaque noeud ;

v. choisir le prédiseur $1^k$ dont la division donne l'incertitude la plus faible et diviser les données de formation en deux noeuds enfants en se basant sur le $1^k$ la division calculée par ordinateur, à chaque noeud enfant étant attribuée une distribution de type suivant les données de formation au noeud enfant;

f) répéter pour chaque noeud enfant si la quantité de données de formation au noeud enfant est supérieure à un seuil ;

g) entrer une expression à reconnaître ;

h) convertir l'expression en un signal électrique ;

i) représenter le signal électrique par une série de vecteurs de caractéristiques quantifiées;

j) comparer la série des vecteurs de caractéristiques quantifiées avec les vecteurs prototypes des caractéristiques enregistrés pour déterminer la concordance la plus proche et assigner une étiquette d'entrée à chacun des vecteurs de caractéristiques de la série correspondant à l'étiquette du vecteur prototype de caractéristique de plus grande concordance;

k) traverser un des arbres de décision pour chaque tranche de temps d'une séquence d'entrée d'une parole pour déterminer une distribution de probabilité pour chaque tranche de temps, la distribution de probabilité étant la probabilité qu'un noeud soit une limite de son ;

l) comparer les probabilités associées aux tranches de temps avec un seuil, pour identifier certain'es tranches de temps en tant que limites entre des sons ;

m) fournir un résultat acoustique pour tous les sons entre chaque paire donnée de limites ;

n) classer les sons sur la base de ce résultat ;

o) sortir un résultat d'identification en réponse au dit résultat.

5. Méthode selon revendication 4 comprenant en outre les phases qui consistent à :

traverser un ou plusieurs d'un deuxième ensemble d'arbres de décision pour chaque tranche de temps sur une séquence de paroles d'entrée pour déterminer une deuxième distribution de probabilité, la distribution de probabilité étant une distribution au-dessus de tous les rangs possibles qu' un son peut prendre pour obtenir un classement parmi les pires d'un son correctement identifié en choisissant le pire classement comme valeur de classe à laquelle distribution de probabilité cumulative des classes dépasse un seuil précis ;

attribuer comme le pire classement absolu parmi les pires classements entre deux limites adjacentes de son, le pire classement du son correctement identifié, entre les limites de son; rejeter toutes les limites de son dont le rang est plus mauvais que ce pire classement absolu dans le segment courant;

faire une liste courte pour le segment;

sortir un résultat d'identification en réponse à la liste courte.

6. Appareil pour la reconnaissance de la parole, comportant:

a) un moyen pour entrer une pluralité de mots des données de formation ;

b) un moyen pour former un ou plusieurs premiers arbres binaires de décision pour poser la question la plus instructive à chaque noeud, en se basant sur l'information contextuelle dans les données de formation, où chaque premier arbre binaire de décision peut correspondre à un temps différent dans une séquence des données de formation;

c) un moyen pour traverser un des arbres de décision pour chaque tranche de temps d'une séquence d'entrée du discours pour déterminer une distribution de probabilité pour chaque tranche de temps, la distribution de probabilité étant la probabilité qu'un noeud soit une limite de son ;

d) un moyen pour comparer les probabilités associées aux tranches de temps à un seuil pour identifier quelques tranches de temps comme limites de sons ;

e) un moyen pour fournir un résultat acoustique pour toutes les limites de temps entre chaque paire donnée de limites ;

f) un moyen pour classer les sons sur la base de ce résultat;

g) un moyen pour sortir un résultat de reconnaissance en réponse au résultat.

7. Appareil selon la revendication 6 comprenant en outre :

h) un moyen pour parcourir un ou plusieurs d'un deuxième ensemble d'arbres de décision pour chaque tranche de temps sur une séquence d'entrée du discours pour déterminer une deuxième distribution de probabilité, la probabilité de distribution étant une distribution au-dessus de tous les rangs possibles que le son correct peut prendre pour obtenir un plus classement parmi les pires d'un son correctement identifié en choisissant le pire classement comme valeur type à laquelle la distribution de probabilité cumulative des classes dépasse un seuil précis ;

i) un moyen pour attribuer comme le pire classement possible parmi les classements les pires entre deux limites de son adjacentes quelconques, le pire classement du son correctement identifié entre les limites du son ;

j) un moyen pour négliger tous les sons dont le rang est plus mauvais que ce pire classement absolu dans le segment courant ;

k) un moyen pour faire une courte liste des sons pour le segment;

l) un moyen pour sortir un résultat d'identification si la liste courte du résultat d'identification étant une liste courte de mots.

8. Appareil selon la revendication 6, comprenant en outre :

h) un moyen pour traverser un ou plusieurs arbres d'un deuxième ensemble d'arbres de décision pour chaque tranche de temps sur une séquence d'entrée du discours pour déterminer une deuxième distribution de probabilité, la distribution de probabilité étant une distribution sur tous les rangs possibles qu'un son peut prendre pour obtenir un pire classement d'un son correctement identifié en choisissant le pire classement comme valeur type à laquelle la distribution de probabilité cumulative des types dépasse un seuil précis ;

i) un moyen pour attribuer en tant que pire classement possible de tous les pires classements entre deux limites de sons adjacentes quelconques, le pire classement du son correctement reconnu entre les limites du son ;

j) un moyen pour négliger tous les sons dont le rang est plus mauvais que ce pire classement absolu dans le segment courant ;

k) un moyen pour faire une courte liste des sons pour le segment;

l) un moyen pour comparer les sons constituant un mot avec un vocabulaire pour voir si le mot se trouve dans la liste courte et établir une courte liste de mots ;

m) un moyen pour sortir un résultat de reconnaissance en comparant les mots de la liste courte avec un modèle de langue pour déterminer la concordance la plus probable d'un mot pour la séquence de parole entrée.

9. Appareil de reconnaissance de la parole, comprenant :

a) un moyen pour entrer un suite d'expressions constituant des données de formation;

b) un moyen pour convertir les expressions des données de formation en signaux électriques ;

c) un moyen pour représenter le signal électrique des données de formation par des vecteurs prototypes, un vecteur représentant une tranche de temps donnée ;

d) un moyen pour attribuer à chaque vecteur de caractéristique prototype, une étiquette de classe associée au vecteur prototype de caractéristique quantifiée ;

e) un moyen pour former un ou plusieurs arbres de décision binaires pour différents temps dans les données de formation, chaque arbre ayant un noeud racine et une pluralité de noeuds enfants, comprenant les phases suivantes :

i) un moyen pour créer un ensemble d'enregistrements de formation comportant 2K+1 les prédiseurs, $1^k$, et une classe prévue, p, où les prédiseurs 2K+1 sont des étiquettes de vecteur de dispositif 2K+1 aux temps consécutifs t-K, ..., t,..., t,..., :t+K, et la classe prévue est un indicateur d'enregistrement binaire si le temps t est associé à une limite de son dans le cas du premier arbre de décision ou est associé au son correct dans le cas du deuxième arbre de décision;

ii. un moyen pour calculer la distribution commune estimée des prédiseurs $1^k$ et du son p pour les prédiseurs 2K+1 en utilisant les données de formation, où les prédiseurs sont les étiquettes de vecteur de caractéristiques aux temps t-K,..., t,..., t+K et p est le son au temps t;

iii. un moyen pour enregistrer la distribution commune estimée de $1^k$ et de p et une distribution correspondante pour chaque prédiseur $1^k$ au noeud racine;

iv. un moyen pour calculer la meilleure division des valeurs que le prédiseur $1^k$ peut prendre pour chaque $1^k$ pour minimiser l'incertitude sur le son à chaque noeud ;

v. un moyen pour choisir le prédiseur $1^k$ dont la division donne l'incertitude la plus faible et diviser les données de formation en deux noeuds enfants en se basant sur le $1^k$ la division calculée par ordinateur, à chaque noeud enfant étant attribuée une distribution de type suivant les données de formation au noeud

enfant;

f) un moyen pour répéter pour chaque noeud enfant si la quantité de données de formation au noeud enfant est supérieure à un seuil ;

g) un moyen pour entrer une expression à reconnaître ;

h) un moyen pour convertir l'expression en un signal électrique ;

i) un moyen pour représenter le signal électrique par une série de vecteurs de caractéristiques quantifiées ;

j) un moyen pour comparer la série des vecteurs de caractéristiques quantifiées avec les vecteurs prototypes des caractéristiques enregistrés pour déterminer la concordance la plus proche et assigner une étiquette d'entrée à chacun des vecteurs de caractéristiques de la série correspondant à l'étiquette du vecteur prototype de caractéristique de plus grande concordance ;

k) un moyen pour traverser un des arbres de décision pour chaque tranche de temps d'une séquence d'entrée d'une parole pour déterminer une distribution de probabilité pour chaque tranche de temps, la distribution de probabilité étant la probabilité qu'un noeud soit une limite de son ;

l) un moyen pour comparer les probabilités associées aux tranches de temps avec un seuil, pour identifier certain'es tranches de temps en tant que limites entre des sons;

m) un moyen pour fournir un résultat acoustique pour tous les sons entre chaque paire donnée de limites ;

n) un moyen pour classer les sons sur la base de ce résultat ;

o) un moyen pour sortir un résultat de reconnaissance en réponse au dit résultat.

**10.** Appareil selon revendication 9 comprenant en outre :

un moyen pour traverser un ou plusieurs d'un deuxième ensemble d'arbres de décision pour chaque tranche de temps sur une séquence de paroles d'entrée pour déterminer une deuxième distribution de probabilité, la distribution de probabilité étant une distribution au-dessus de tous les rangs possibles qu' un son peut prendre pour obtenir un classement parmi les pires d'un son correctement identifié en choisissant le pire classement comme valeur de classe à laquelle distribution de probabilité cumulative des classes dépasse un seuil précis ;

un moyen pour attribuer comme le pire classement absolu parmi les pires classements entre deux limites adjacentes de son, le pire classement du son correctement identifié, entre les limites de son;

un moyen pour ignorer toutes les limites de son dont le rang est plus mauvais que ce pire classement absolu dans le segment courant;

un moyen pour faire une liste courte pour le segment;

un moyen pour sortir un résultat d'identification en réponse à la liste courte.

FIG. 1.

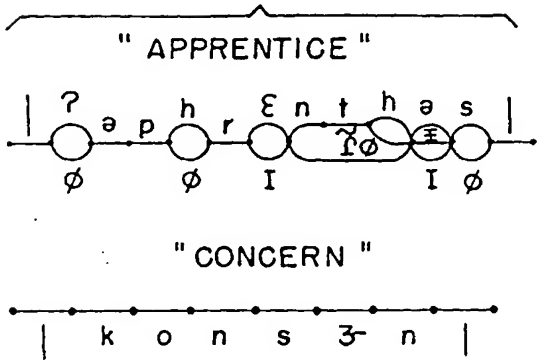



FIG. 2.

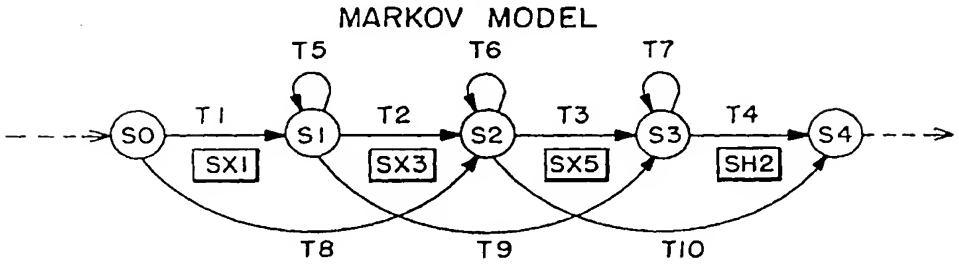



FIG. 3.

| PHONE | STATE | TRANS-ITION | ARC PROB. | OUTPUT PROBABILITIES L001 L002 ... L200 | WEIGHT |
|---|---|---|---|---|---|
| PH1 | S10 | T101 | 0.75 | 0.03 0.00 ..... 0.13 | W(1) |
| | | T102 | 0.25 | | |
| | S11 | T111 | 0.50 | | |
| | | T112 | 0.25 | | |
| | | T113 | 0.25 | | |
| | S12 | T121 | | | |
| | | T122 | | | |
| | | T123 | | | |
| | S13 | T131 | | | |
| | | T132 | | | |
| PH2 | S20 | T201 | | | W(2) |
| | | T202 | | | |
| PHN | SN0 | TN01 | | | W(N) |
| | | TN02 | | | |
| | SN1 | TN11 | | | |
| | | TN12 | | | |
| | | TN13 | | | |
| | SN2 | TN21 | 0.3 | | |
| | | TN22 | 0.7 | 0.25 0.12 ..... 0.00 | |
| STATISTICAL WORD MODEL | | | | | W(ø) |

FIG. 4.

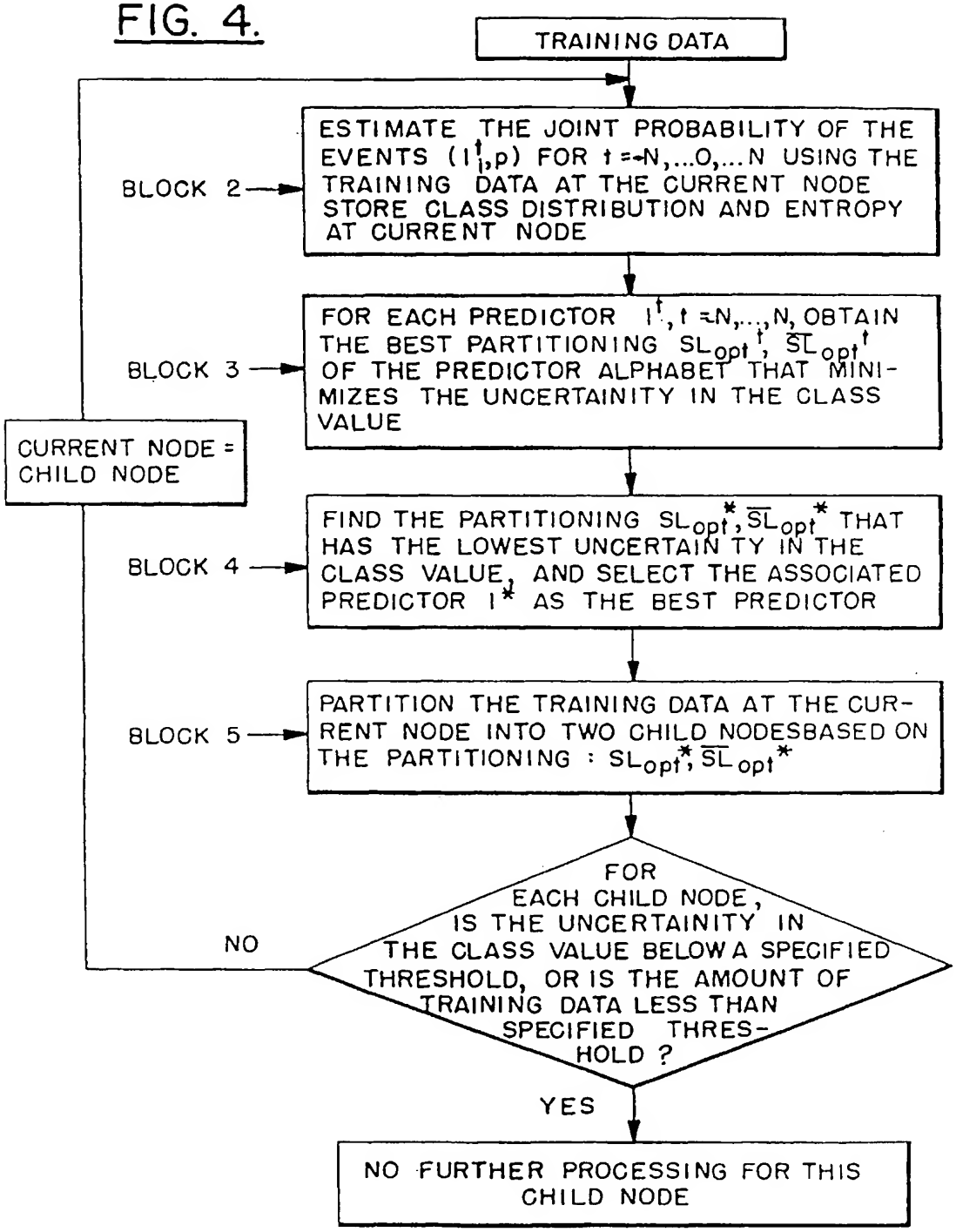

TRAINING DATA
BLOCK 2
ESTIMATE THE JOINT PROBABILITY OF THE EVENTS $(I_j^t, p)$ FOR $t = -N, \ldots 0, \ldots N$ USING THE TRAINING DATA AT THE CURRENT NODE STORE CLASS DISTRIBUTION AND ENTROPY AT CURRENT NODE
BLOCK 3
FOR EACH PREDICTOR $I^t$, $t = -N, \ldots, N$, OBTAIN THE BEST PARTITIONING $SL_{opt}^t$, $\overline{SL}_{opt}^t$ OF THE PREDICTOR ALPHABET THAT MINIMIZES THE UNCERTAINITY IN THE CLASS VALUE
CURRENT NODE = CHILD NODE
BLOCK 4
FIND THE PARTITIONING $SL_{opt}^*, \overline{SL}_{opt}^*$ THAT HAS THE LOWEST UNCERTAIN TY IN THE CLASS VALUE, AND SELECT THE ASSOCIATED PREDICTOR $I^*$ AS THE BEST PREDICTOR
BLOCK 5
PARTITION THE TRAINING DATA AT THE CURRENT NODE INTO TWO CHILD NODESBASED ON THE PARTITIONING : $SL_{opt}^*, \overline{SL}_{opt}^*$
FOR EACH CHILD NODE, IS THE UNCERTAINITY IN THE CLASS VALUE BELOW A SPECIFIED THRESHOLD, OR IS THE AMOUNT OF TRAINING DATA LESS THAN SPECIFIED THRESHOLD ?
NO
YES
NO FURTHER PROCESSING FOR THIS CHILD NODE

FIG. 5.

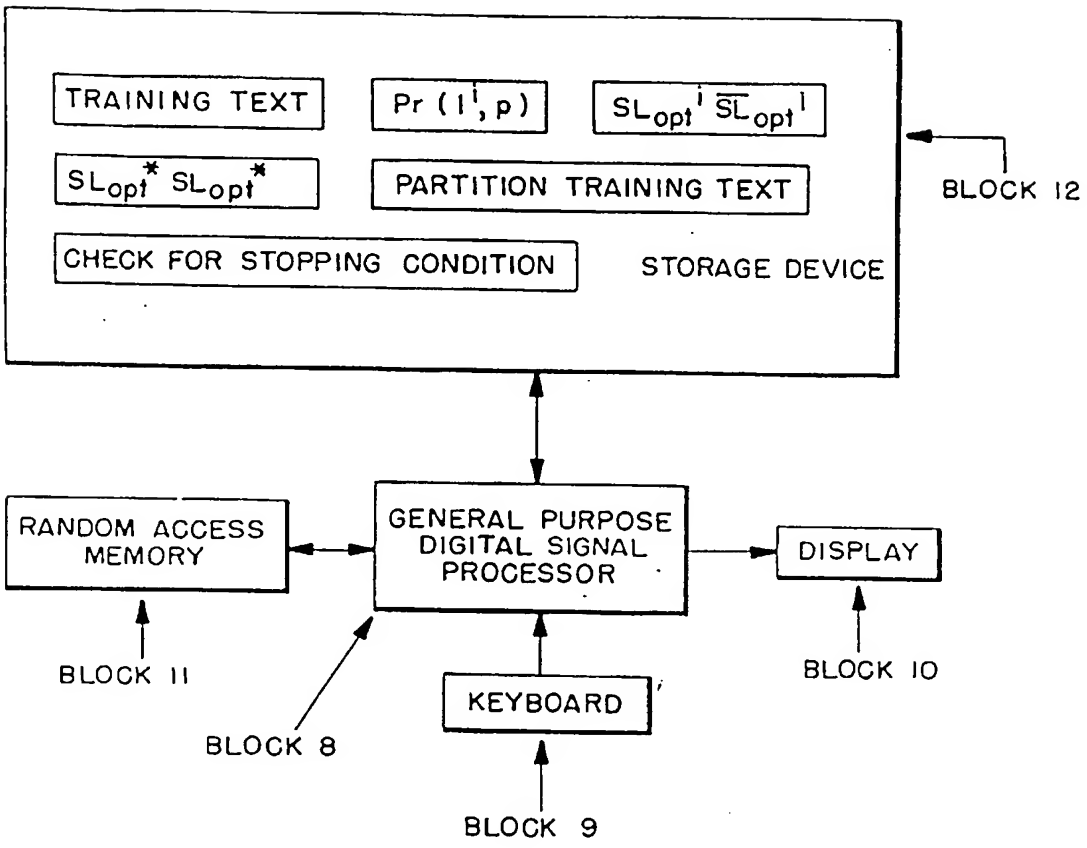

TRAINING TEXT
Pr (l^i, p)
SL_opt^i SL̄_opt^i
SL_opt* SL_opt*
PARTITION TRAINING TEXT
CHECK FOR STOPPING CONDITION
STORAGE DEVICE
BLOCK 12
RANDOM ACCESS MEMORY
GENERAL PURPOSE DIGITAL SIGNAL PROCESSOR
DISPLAY
BLOCK 11
BLOCK 8
KEYBOARD
BLOCK 10
BLOCK 9

FIG. 6.

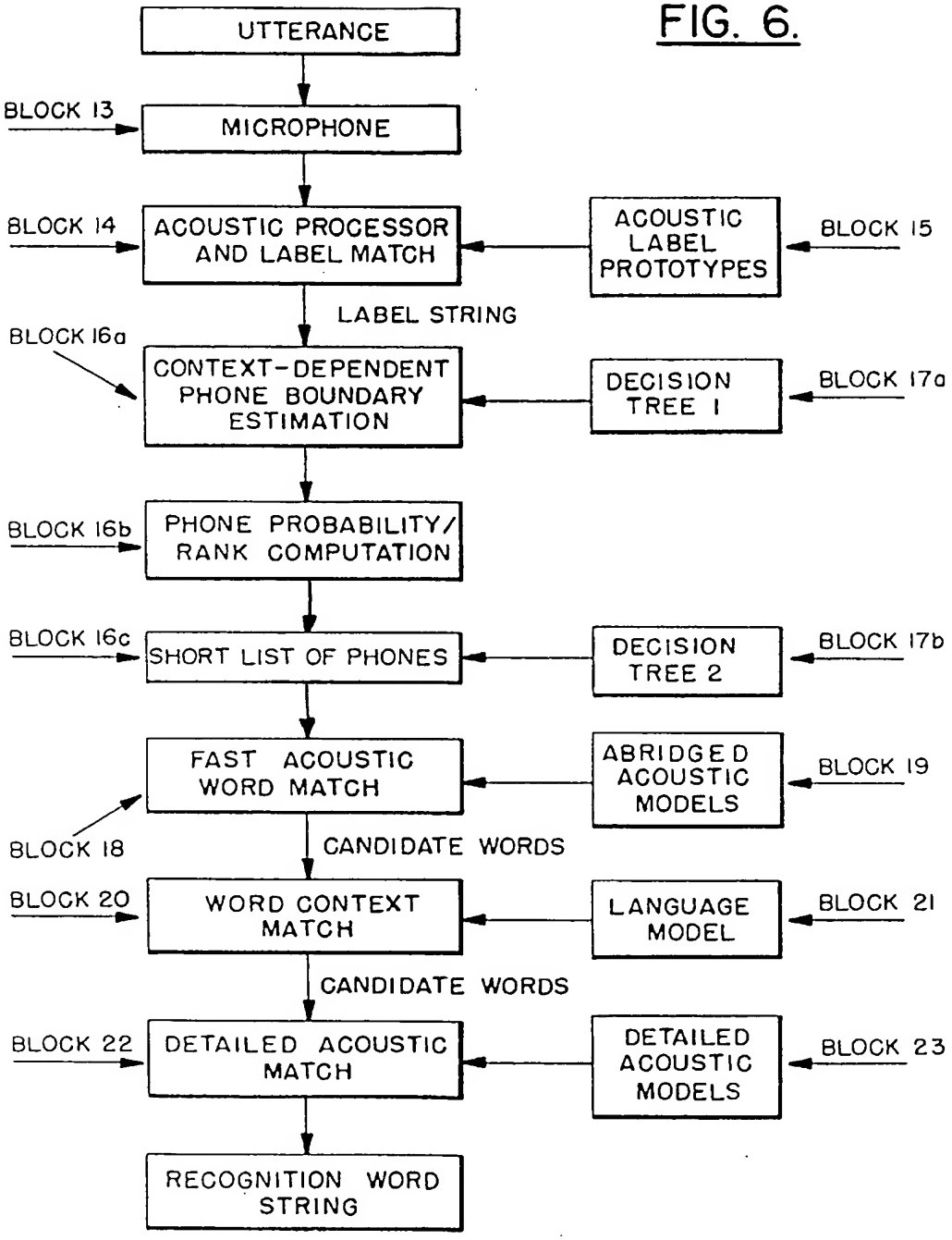
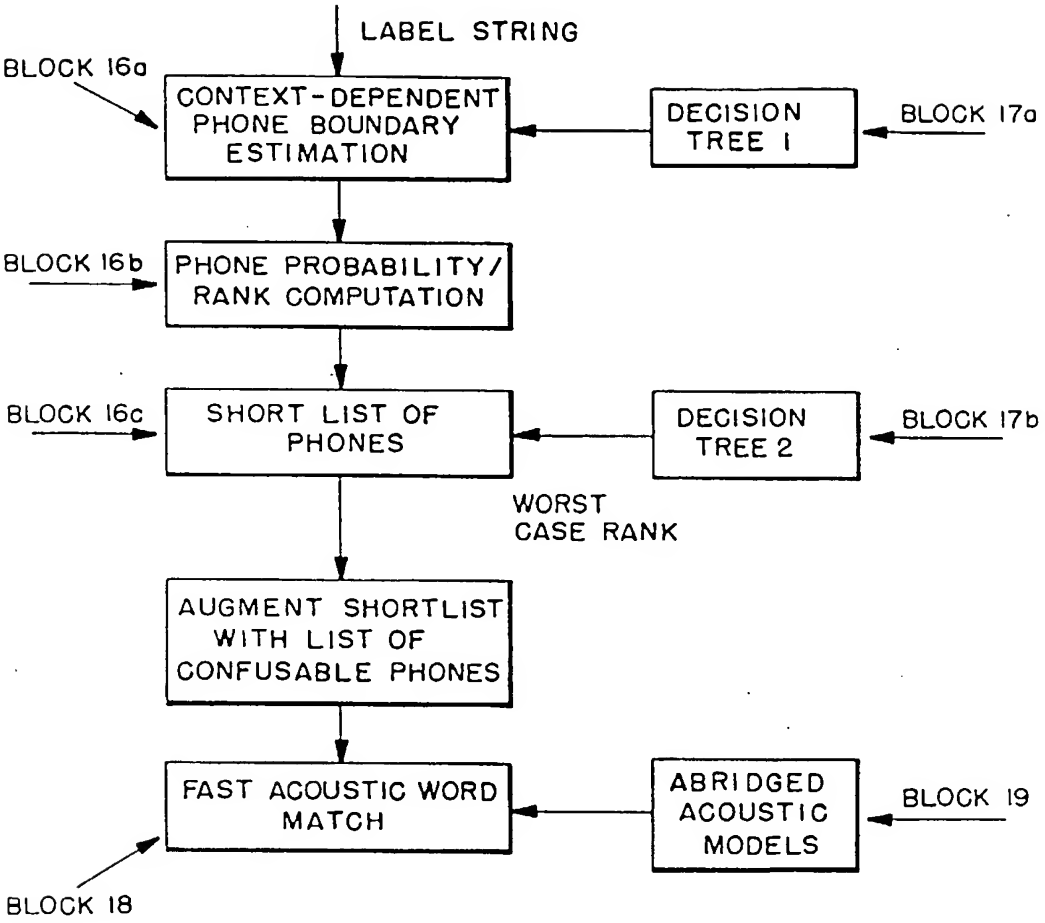

UTTERANCE
BLOCK 13
MICROPHONE
BLOCK 14
ACOUSTIC PROCESSOR AND LABEL MATCH
ACOUSTIC LABEL PROTOTYPES
BLOCK 15
LABEL STRING
BLOCK 16a
CONTEXT-DEPENDENT PHONE BOUNDARY ESTIMATION
DECISION TREE 1
BLOCK 17a
BLOCK 16b
PHONE PROBABILITY/ RANK COMPUTATION
BLOCK 16c
SHORT LIST OF PHONES
DECISION TREE 2
BLOCK 17b
FAST ACOUSTIC WORD MATCH
ABRIDGED ACOUSTIC MODELS
BLOCK 19
BLOCK 18
CANDIDATE WORDS
BLOCK 20
WORD CONTEXT MATCH
LANGUAGE MODEL
BLOCK 21
CANDIDATE WORDS
BLOCK 22
DETAILED ACOUSTIC MATCH
DETAILED ACOUSTIC MODELS
BLOCK 23
RECOGNITION WORD STRING

FIG. 7.


LABEL STRING
BLOCK 16a
CONTEXT-DEPENDENT PHONE BOUNDARY ESTIMATION
DECISION TREE 1
BLOCK 17a
BLOCK 16b
PHONE PROBABILITY/ RANK COMPUTATION
BLOCK 16c
SHORT LIST OF PHONES
DECISION TREE 2
BLOCK 17b
WORST CASE RANK
AUGMENT SHORTLIST WITH LIST OF CONFUSABLE PHONES
FAST ACOUSTIC WORD MATCH
ABRIDGED ACOUSTIC MODELS
BLOCK 19
BLOCK 18